





Ex LIBRIS  
UNIVERSITATIS  
ALBERTAENSIS














Digitized by the Internet Archive  
in 2024 with funding from  
University of Alberta Library

<https://archive.org/details/Ingram1973>









THE UNIVERSITY OF ALBERTA

RELEASE FORM

NAME OF AUTHOR ..... WAYNE D. INGRAM .....

TITLE OF THESIS ..... COMPUTER RECOGNITION OF  
..... PATTERNS IN SCIENTIFIC AND  
..... TECHNICAL WRITING .....

DEGREE FOR WHICH THESIS WAS PRESENTED ..... M-Sc. ....

YEAR THIS DEGREE GRANTED ..... 1973 .....

Permission is hereby granted to THE UNIVERSITY OF  
ALBERTA LIBRARY to reproduce single copies of this  
thesis and to lend or sell such copies for private,  
scholarly or scientific research purposes only.

The author reserves other publication rights, and  
neither the thesis nor extensive extracts from it may  
be printed or otherwise reproduced without the author's  
written permission.





THE UNIVERSITY OF ALBERTA

COMPUTER RECOGNITION OF PATTERNS

IN

SCIENTIFIC AND TECHNICAL WRITING

by



WAYNE D. INGRAM

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH  
IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE  
DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF COMPUTING SCIENCE

SPRING, 1973





THE UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled COMPUTER RECOGNITION OF PATTERNS IN SCIENTIFIC AND TECHNICAL WRITING submitted by Wayne D. Ingram in partial fulfilment of the requirement for the degree of Master of Science.



## ABSTRACT

Linking of various types of users with suitable data banks will become a problem with the growth of information networks. The documents in the data banks display characteristic styles; these styles are a product of content and form, as imposed by subject, author, and audience. Attention is directed to the style of scientific and technical documents, as these form the bases of many present day data banks. This thesis suggests a theory of discourse suited to the representation of such styles through the graphic specification of the intersentence structure of the discourse. A system is described whereby certain structural characteristics of style are recognized by the computer and are assembled into a pattern output on a CalComp plotter. Samples from various levels of scientific and technical writing are analyzed, and present and future applications of the model are discussed.





## ACKNOWLEDGEMENTS

I would like to express my deep and sincere appreciation to my thesis supervisor, Doreen M. Heaps, for her many contributions, her advice and criticism, and her unfailing support throughout the duration of this research.

I would also like to express my gratitude to Mary Lois Marckworth of the Department of Linguistics, University of Alberta, for her criticism and suggestions during the formative period of the project.

And, I would like to thank my consultant on all matters technical, Leslie-Ann Ingram.

The financial assistance of the National Research Council of Canada is gratefully acknowledged.



## TABLE OF CONTENTS

	Page
CHAPTER I - INTRODUCTION . . . . .	1
1.1 The Demand for Information . . . . .	1
1.2 The Objective and Approach . . . . .	11
CHAPTER II - THE STRUCTURE OF SCIENTIFIC AND TECHNICAL DISCOURSE: A PROPOSED THEORY . . . . .	17
2.1 Introduction and Discussion of the Theory . .	17
2.2 The Sentence Outline . . . . .	21
2.3 Recognition of Structure . . . . .	23
2.3.1 Dependency . . . . .	27
2.3.2 Content Relations . . . . .	28
2.3.3 Pattern Assembly . . . . .	30
CHAPTER III - IMPLEMENTATION OF THE MODEL . . . . .	35
3.1 Introduction . . . . .	35
3.2 General Procedure for Unedited Text . . . . .	36
3.3 Programming Description . . . . .	40
3.3.1 PLATEXT Main Program . . . . .	40
3.3.1.1 BLOCK DATA . . . . .	40
3.3.1.2 Dimensioning and Initializa- tion . . . . .	41
3.3.1.3 Input Files . . . . .	47
3.3.1.4 Sentence Boundary Definition	49
3.3.1.5 Word Boundary Definition . .	53
3.3.1.6 Matching of Dictionary Entries and Content Words . . . . .	57
3.3.1.7 Dominant Attributes . . . . .	63





	Page
3.3.1.8 Pattern Generation . . . . .	65
3.3.1.9 Example . . . . .	74
3.3.2 Subroutines: Subroutine BEFIX . . .	78
3.3.2.1 Subroutine RECNUM (A, B, C)	93
3.3.2.2 SUBROUTINE LINSYM (A, B, C, D) . . . . .	94
3.3.2.3 Subroutines SUGHB1, SUGHB2, AND SUGHB3 . . . . .	94
3.3.3 CalComp Library Subroutines Used in PLATEXT . . . . .	96
3.3.4 Library Subroutine ICMPAR . . . . .	98
3.4 General Procedure for Augmented Text . . . .	100
CHAPTER IV - SAMPLE TEXT . . . . .	105
4.1 Extract from <u>The Genetic Code</u> . . . . .	105
4.2 The Corpus of Samples . . . . .	116
4.2.1 Introduction . . . . .	116
4.2.2 Category I; Samples 4, 6, 7, 8, 10 .	117
4.2.3 Category II: Samples 1, 11, 14 . . .	118
4.2.4 Category III: Samples 2, 3, 5, 9, 12, 13, 15 . . . . .	119
CHAPTER V BACKGROUND STUDIES . . . . .	122
5.1 Introduction . . . . .	122
5.2 Harris' "Discourse Analysis" . . . . .	123
5.3 Relation to PLATEXT . . . . .	128
5.4 Jacobson's Sentence-Connecting Routine . . .	130
5.5 Relation to PLATEXT . . . . .	141
5.6 Related Fields of Study . . . . .	144
5.6.1 Introduction . . . . .	144



	vii
	Page
5.6.2 Text Handling Methods Based on Statistics . . . . .	145
5.6.2.1 Bibliometrics . . . . .	145
5.6.2.2 Statistical Inference . . . . .	149
5.6.3 Content Analysis . . . . .	151
5.6.4 Auto-Abstracting . . . . .	157
5.6.5 General References . . . . .	158
CHAPTER VI CONCLUDING DISCUSSION. . . . .	159
6.1 General Comments and Immediate Relevance. . . . .	159
6.1.1 Automatic Abstracting. . . . .	159
6.1.2 Content Analysis . . . . .	161
6.1.3 Style Analysis . . . . .	162
6.2 Immediate Development . . . . .	163
6.3 Two Steps Past PLATEXT . . . . .	165
BIBLIOGRAPHY . . . . .	169
APPENDIX A - DICTIONARY OF RELATION INDICATORS . . . . .	178
APPENDIX B - ERASURE LIST. . . . .	182
APPENDIX C - SAMPLE DIAGRAMS C-1 TO C-15 . . . . .	186
APPENDIX D - SAMPLE TEXTS 1 TO 15. . . . .	202
APPENDIX E - SAMPLE TEXT (AFTER JACOBSON). . . . .	203
APPENDIX F - PLATEXT SOURCE LISTING. . . . .	206





## LIST OF FIGURES

	Page
Fig. 1: Levels in the Scientific and Technological Community . . . . .	7
Fig. 2: Data Bank and User Matching . . . . .	12
Fig. 3: A Hypothetical Sentence Outline . . . . .	22
Fig. 4: Sentence Outline with Additions . . . . .	24
Fig. 5: Sentence Dependency Hierarchy . . . . .	26
Fig. 6: Content Linking . . . . .	29
Fig. 7a: Sentence Outline . . . . .	32
7b: Sentence, or Text, Diagram . . . . .	33
7c: Sentence, or Text, Diagram . . . . .	34
Fig. 8: Sequence of Text Operations . . . . .	38
Fig. 9: Sentence Isolation Routine . . . . .	50
Fig. 10a: Sample Input I . . . . .	52
10b: Sample Input II . . . . .	52
Fig. 11: Word Isolation Routine . . . . .	54
Fig. 12: Parentheses; Two Uses . . . . .	55
Fig. 13: Hyphen and/or Short Dash; Five Uses . . . . .	55
Fig. 14: Dictionary Matching Routine . . . . .	58
Fig. 15: Erasure List and Content Word Matching Routine . . . . .	59
Fig. 16: Establishing the Dominant Attributes of the Sentence . . . . .	64
Fig. 17: Table of Sentence Characteristics . . . . .	67
Fig. 18: Pattern Synthesis . . . . .	69
Fig. 19: Internal Representation . . . . .	73



	ix
	Page
Fig. 20: Plot Production Routine . . . . .	75
Fig. 21: Sample; Original Order, with Dictionary and Content Tagging . . . . .	77
Fig. 22: Internal Representation. . . . .	79
Fig. 23: Pattern, and Sample in Revised Order . . . .	80
Fig. 24: Subroutine Linkage . . . . .	81
Fig. 25: Three Possible Linkages Between Sentence Symbol 2 and Sentence Symbol 6 . . . . .	85
Fig. 26: Sentence Symbol Assembly . . . . .	87
Fig. 27: Subroutine BEFIX; Superordinate and Top-Level Coordinate Linkage . . . . .	88
Fig. 28: Subroutine BEFIX; Subordinate Linkage. . . .	89
Fig. 29: Subroutine BEFIX; Coordinate Linkage . . . .	91
Fig. 30: Rectangle; Orientation . . . . .	95
Fig. 31: Example; CALL NUMBER (A, B, C, NUMB (3), E, 2) . . . . .	99
Fig. 32: Rotation Parameter . . . . .	99
Fig. 33: Sample of Augmented Text . . . . .	101
Fig. 34: Sentences 1-3. . . . .	106
Fig. 35: Sentences 4-9. . . . .	107
Fig. 36: Sentences 10-15. . . . .	108
Fig. 37: Sentences 16-21. . . . .	109
Fig. 38: Sentences 22-25. . . . .	110
Fig. 39: Sentences 26-31. . . . .	111
Fig. 40: Text Diagram; Sample from <u>The Genetic Code</u> .	112
Fig. 41: Text Summary; Sentences from Top Level of Figure 40. . . . .	112
Fig. 42a: Sample Advertisement (after Harris). . . . .	126
42b: Double Array (after Harris). . . . .	126





Fig. 43:	Sentence Dependency Hierarchy (after Jacobson). . . . .	131
Fig. 44:	Routing Procedure I (after Jacobson) . . . .	134
Fig. 45:	Routing Procedure II (after Jacobson). . . .	135
Fig. 46:	Routing Procedure III (after Jacobson) . . .	136
Fig. 47:	Symbols (after Jacobson) . . . . .	138
Fig. 48:	Principal Path (after Jacobson). . . . .	139
Fig. 49:	Full Text Diagram (after Jacobson) . . . . .	140
Fig. 50:	MAPTEXT (after Sedelow). . . . .	153
Fig. 51:	VIA (after Sedelow). . . . .	155



## CHAPTER I

### INTRODUCTION

#### 1.1 The Demand for Information

For the mid-twentieth century technological man, one of the facts of life has been the so-called "information explosion", brought about by large scale and rapidly evolving developments in many fields. These include technical advances, which make it possible for publishing houses to produce very large quantities of material in a short time (1), economic conditions that make it both necessary and seemingly profitable to produce such quantities, and social conditions that exert pressure on individuals to publish, sometimes to "publish or perish", and thus to contribute to the output of printed words (2). In the background, and sometimes overlooked as one of the basic causes, is the population explosion, which has created more scientists and technologists, as well as the population base to support them (3).

In the course of the past ten years the information explosion has been perhaps the best documented of all the "explosions", but this has not led to any diminution in the problems created by it. The large and uneven corpus of publications about research on all aspects of information compounds the difficulties for the person who wishes to study the information explosion and its resulting conditions,





confrontations, and solutions. Hence, many of these investigations must begin with a very comprehensive and frequently difficult literature search.

The searcher finds that the literature contains many suggestions that could be regarded as idealized solutions to the problem as a whole but, even yet, very few simple workable approaches that might effect immediate improvements. For example, in many of the discussions concerning the attainment of long-range goals we find plans for information networks, with linked information centres offering on-line retrieval for everyone in the society. Such networks are frequently envisaged as being brought into operation with the aid of complex computers and under the direction of experts in operations research. The emphasis on long-range goals has developed as a primitive science of information management has slowly grown and as it has contributed to knowledge of how information is transferred. The corresponding evolution of large scale computer networks has led to further emphasis on overall goals. In addition, political pressure to make information widely available or to transmute and transform it from one stratum of society to another has contributed to the increasing awareness of the need for the intelligent establishment of national objectives, policies, and administrative structures in the field of information transfer (4).

Fortunately, at the same time, a better understanding of machine limitations and a greater appreciation of human capabilities has resulted in basic research being directed



to paths more suitable than those sometimes followed in the past. An example of this is the shift from research leading to fully automated translation to research into the way a single language conveys information, and into how such processes might relate to a machine (5). As a result of the redirection of research we may perhaps hope for the appearance of more practical solutions to our immediate informational ills. Work done in the last few years seems to bear this out. Improvements have been made in computer supervisory systems, in retrieval languages, and in related aspects of computer software. These advances have been supplemented by more efficient storage and search methods. Automated classification, indexing, and dissemination techniques have become more sophisticated. Increased use of the cost study approach has helped to provide accurate assessments of methods, and has occasionally functioned as an unwelcome link with reality.

The most significant research into information handling has been carried out in the field of science and technology; here the proliferation of material has been greatest and the demand for relevant material has been most acute (6). Despite recent emphasis on the need for information from other fields, it is probable that the transfer of scientific and technical information will continue to occupy much attention and to pose very important problems. Contemporary North American society is very much the child of science and technology, albeit a sometimes unwilling child. Given the acceleration of change in our society, it





is probably true that the demand for scientific and technical information will not diminish, although it may take very different forms and require different emphasis.

Scientific and technical information has several crucial characteristics. Initially, the material almost always appears in relatively short papers; these papers are submitted to and published by one of the large number of learned journals that are to be found in many parts of the world (7, 8). The information must be available quickly, for, because of the cumulative nature of science, most articles are very "timely". The scientist needs and wants only the very relevant papers; otherwise he will be faced with an impossible number of documents to scan. Further problems will arise as the number of documents and scientists continue to increase. Dr. J. Belzer, Professor of Information and Communication Science at the University of Pittsburgh, predicts that the individual scientist or technologist

. . . will concentrate more and more on his own special interests for his reading needs, and this concentration will tend to create a larger number of classes of readers. The precision required to place the literature in these classes will become very great. We are being faced with the problem of providing greater precision than before. Unfortunately, the information retrieval systems of the present day are having precision difficulties, and this problem is getting worse.

This excerpt from a paper (9) delivered to a 1969 symposium on Education and Information Science indicates a number of trends which has been developing as the volume of printed scientific and technical literature in any given field



has passed well beyond the point where a single scientist can scan it in its entirety. Most obviously, scientists become specialists and as time goes on and as the literature grows through their efforts and those of their colleagues, each scientist specializes a little more to reduce the amount of relevant material he must scan to be "currently aware". This development will place a great strain, as Dr. Belzer emphasized, on the retrieval system which supplies the specialist with information.

In actuality, the situation is a great deal more complex than Dr. Belzer has painted it. What he has described is a system in which scientists communicate with scientists in a single field; this is a very small part of the processes involved in the communication of scientific and technological information. As has been implied, it is now widely recognized that science and technology, for good or ill, affect the entire community; and scientific and technical information that is correct, relevant, and in a suitable form must be transmitted to many types of users, which may include people as diverse as "small" industrialists and national politicians. In order for wise decisions to be made in many fields, a satisfactory information transfer concerning science and technology must occur throughout all levels of the society (10). An example of both the need for and the concern with the transfer of such information across levels is the Report of the Senate Committee on Science Policy (the Lamontagne Report) (11). Typically, the



levels might be characterized as in Figure 1.

The communication gaps between each of the blocks in Figure 1 arise out of differences inherent in the normal manner of communication characteristic of each group and these in turn are the results of the differing aims and objectives, backgrounds and resources, and written and unwritten traditions of each group. Expressed in simple terms, the differences in the techniques that the members of each group employ to convey information may be said to be a function of their individual styles; different styles practised by different groups form very real barriers to effective communication.

Results of these barriers are the "precision difficulties" mentioned by Dr. Belzer. A scientist in the top-most block of Figure 1 has a certain set of sources (journals, abstract journals, indexes, etc.--probably quite large in number) which he regularly scans as part of his routine. If a reference is made to a paper which seems to be immediately relevant, the scientist endeavours to obtain a copy of it. If the scientist suddenly finds need of information on a newly-related field, he consults his indexes and retrieval system again, and seeks to obtain the necessary information. In each instance, the scientist has a definite requirement in terms of content; what must also be recognized is that he has also a fairly rigid requirement as to form. He does not expect to find useful information for his field of study in the columns of daily newspapers. A retrieval system which





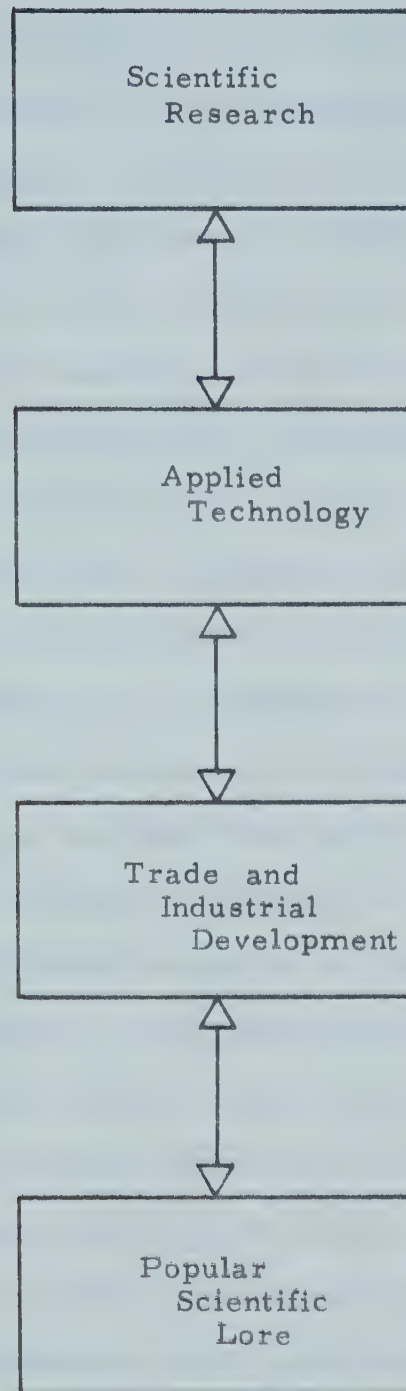


Figure 1: Levels in the Scientific  
and Technological Community



fails to satisfy his demands for both appropriate content and form might be said to be experiencing "precision difficulties". Similar comments can be made concerning other users on other levels. An industrialist is more likely to keep "currently aware" through the scanning of trade journals than by the perusal of pure physics periodicals (12).

The scientist has very legitimate reasons for wanting the content in a certain form; he has need of particular information and his time, at least until very recently, has been genuinely too valuable to waste wading through extraneous material. The scientific literature has developed certain forms and conventions to satisfy him quickly. The same can also be said of the engineer who wants much the same content but in a different form. His point of view is different, and he expects different aspects of the information to be emphasized. A general retrieval system which both scientists and engineers use but which cannot satisfy both is certainly not a successful retrieval system. In times past this problem of precision was not regarded as critical because retrieval systems usually applied only to highly related collections of documents, such as a single journal, or to a small area of research with half a dozen principal sources. More significantly, the system usually catered to a particular group and discipline, and only the needs of these groups were considered important. However, it is now realized that entire new classes of users are appearing to make demands upon retrieval systems and these users may come to be





regarded as equally important by designers and funders. Precision difficulties must increase.

It is only too evident, especially from public discussion and from reports such as the Lamontagne Report, that individual retrieval systems or retrieval networks are coming to be regarded as the answer to all problems of information transfer. They are thought of as the means "to get government to the people",\* as a means to make unsophisticated industry sophisticated,\* and as a source of legal aid for the unlettered\* (13, 14, 15, 16, 17). The problems, precision difficulties included, of information transfer, of information identification, and of information transformation are going to be still further intensified. Recent references in the literature point to concern with this essentially multidisciplinary problem, especially in regard to interchangeable indexing languages, intermediate lexions, and other types of switching points in information systems (18, 19, 20, 21, 22, 23, 24). Over the past three years, at the University of Alberta, certain aspects of this problem have been investigated by Mercier and others (25, 26, 27, 28, 29).

It is evident that, in the future, the successful large retrieval system or network must consist of a set of

---

\*Cf. Political campaigns that promise to "get government to the people" are then faced with the problems of implementation; an instance is the Alberta "NOW" campaign (13, 14). International conferences have as themes "information services for underdeveloped countries" (15, 16). Government grants are given to investigate the efficiency of "community information centres" (17).



linked subsystems, each servicing a user group. Each subsystem will have its own data base into which and from which information must be channelled. The decision specifying which documents are associated with which data base has customarily been made by humans, who recognize both the content and the characteristics of style that mark a particular document as meant for a certain user group. But if the demand for multipurpose retrieval systems, of the type discussed, does increase, the supply of human identifiers, which is even now not adequate, will be completely insufficient; people who make such decisions need to have certain native characteristics; they require, in addition, fairly lengthy training; they also seem to be very necessary, as has been indicated by user dissatisfaction with retrieval results from data bases not adequately screened by competent personnel. To sum up, they are both expensive and scarce.

If some technique could be devised to replace or supplement these "large scale identifiers"--they are not indexers, working within a subgroup of knowledge--this would be at least one step forward in our attempts to make large scale systems effective. It has been indicated already that certain tools, such as intermediate lexicons, provide some relief; they aid switching once placement identification has been made. But there is little evidence of any systematic approach to the problem of this initial categorization and classification.



## 1.2 The Objective and Approach

This thesis will be concerned with the placement aspect of the overall problem posed by the demand for large scale networks. It will describe the identification, by automated means, of various levels of technical and scientific style. As already discussed, it is assumed that these levels of styles, the style arising from form plus content (30), are indicative of groups of information users and information generators. It is postulated that, if such groups of styles can be automatically found in documents, then these documents may be automatically assigned to an appropriate data base within an overall system (see Figure 2). The analysis is confined to the field of scientific and technical writing. Some justification for this restriction has already been given in this introductory section. Further discussion of the point will be found in Chapter 2.

Significant implications arise if such a method can be successfully implemented and these are discussed in the conclusions. In this introductory section the approach of the thesis is outlined very briefly. It is hoped that the description of the method, given in the succeeding chapters, will make evident to the reader all the details of the underlying assumptions.

The primary effort of the thesis has gone into the development of a method for computer recognition and representation of intersentence structure, which is made evident through graphic illustration. The graphic





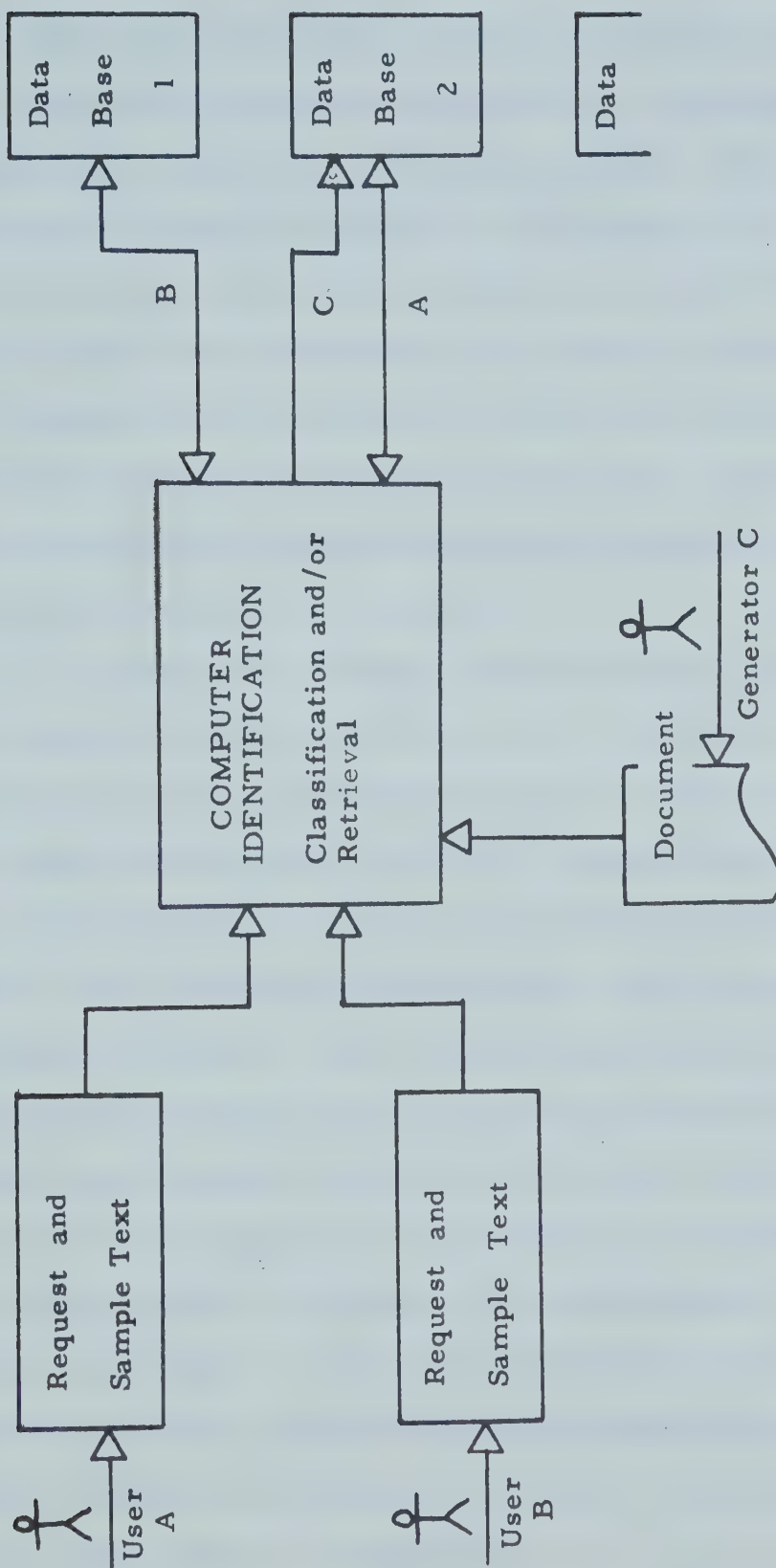


Figure 2: Data Bank and User Matching



representation makes the stylistic differences between levels of writing very evident and serves to complement the already established statistical differentiators. The method presumes that intersentence relations in the text are indicators of style and that the relations can be described in terms of subordinate and coordinate links which, when combined, result in a pattern; this pattern has been ascertained essentially from the surface structure of the text. The components of the analyzed and pattern-synthetic scheme will be described in detail later in the thesis.

It should be noted that the use of text structure for stylistic differentiation does not imply that the computer traces transformations or in some way investigates the deep structure of the text. As has been stated, what is traced are the relations among the individual sentences of the text, as these are apparent from the analysis of the surface structure. This thesis does not presume the validity of any particular linguistic theory or hypothesis connecting the deep structure with the surface structure, not out of disagreement, but because the structure examined is intersentence as opposed to intrasentence. For example, within the scope of this investigation, the generative and transformational grammars of Noam Chomsky and his group are at most marginally relevant (31, 32). As noted, the objective of the thesis is different.

In addition, although it is also presumed that it is helpful to reproduce the pattern found on a graphical output





device, in this instance a CalComp plotter, it is not asserted that the process employed in the pattern-seeking program has an exact or even approximate psycholinguistic analogue in human beings, or that pattern-production or even pattern-recognition is a conscious process for language-users in general. It is suggested that the manner in which sentences are interrelated in the text contributes strongly to the "overall impression" of the text, which is the stylistic factor that is most important in the context of this investigation, and that this overall impression can be reduced and expressed in a graphical form and thus be more readily apparent.

It should also be emphasized that economic realities were kept in mind during the program design. Thus, relatively unedited text was given preference over manually pre-edited text and the internal complexity of the program was held to a minimum. These procedure decisions will be further discussed in the description of the method.

In view of the fact that trained human beings can classify information easily, often without conscious effort, as being suitable or not suitable in form, an objection may be made that this is a superfluous use of the computer, that it is preferable to let computers perform in the areas in which they excel and let people perform those tasks, such as decision-making, which they do best. However if this advice were followed we might have local maximums of efficiency imbedded in a system that was inefficient overall. We would



find ourselves stopping the machine, delivering the data to the human, having the required task performed, delivering the results to the machine, and proceeding with the processing. For the sake of retaining the information processing entirely within the machine and achieving maximum efficiency it is inevitable that the machine be at times asked to perform some task intrinsically better suited to human performance. This approach is one justification for the investigation of an admittedly imperfect, inefficient, and relatively expensive machine analysis of scientific and technical text.

As has been indicated, one should also consider the next or next-to-next generation of information systems, and their possible configurations. For example, although the use of language is entirely a faculty of humans, in the foreseeable future some aspects of that faculty may be extended to human artifacts. Much of the research presently being done in this field is in the form of question-answering machines (33). Perhaps the prototypical application of computing machines that possess true conversational capability will be as the core of an information system. The great advantage of such a truly conversational machine is that the user need not be trained in the rigours of formatting; he or she need only be a member of the language community. One might therefore expect to find eventually a truly conversational machine imbedded in an information system which services requests from the full spectrum of the



scientific-technical-industrial community. An integral part of such an information system must be the arranging and presenting of the requested information in a form readily assimilated by the user who needs recognizable content and form, else the sophisticated subtleties of the system go for naught, and it will be little more useful than existing systems. Unless we are willing to allow only information specialists to organize and present the retrieved information, the goal of a data processing computer with conversational capabilities must be extended to include the capability to communicate in several styles or modes, perhaps in several media, as required. To accomplish this in the "forseeable future", further emphasis needs to be placed on the necessary analytic and synthetic research and the consequent operational development.





## CHAPTER II

### THE STRUCTURE OF SCIENTIFIC AND TECHNICAL

#### DISCOURSE: A PROPOSED THEORY

##### 2.1 Introduction and Discussion of the Theory

As a first step in the thesis project, it was necessary to have a suitable theory of discourse. This might have been found fully specified in the literature, or suggested by the literature, or completely developed ab initio. By "theory of discourse" in this thesis is meant a consistent model of reality which allows one to make quantitative statements about samples of discourse. Discourse is used in a general sense as it is found in Fairthorne and Coblans (34, 35). Fairthorne has frequently stated that "our field (information science) is the study of human discourse as it is recorded in documents" (36). It was also desired that the theory developed should be as reflective of the structure of discourse as possible, that is, it should be qualitative. The frame of reference, the structure of discourse as a whole, meant that the aspects of text structure under consideration would be the intersentence, as opposed to intrasentence, relationships. It should be noted again that the majority of models discussed in the literature of modern linguistics are concerned with intrasentence relationships, and so, as has been indicated, are of marginal relevance.

The fundamental concept governing the theory of



discourse developed for this project is the idea that "levels of relevance" are the primary expression of structure in all discourse, and that these levels are especially evident in certain types of recorded discourse. As an illustration, for the composer, "levels of relevance" might translate to something like "levels of thematic emphasis", and a similar translation is evidenced in the work of graphic artists. For at least one contemporary subculture, "levels of relevance" translates freely to "Levels of Consciousness". The concept of "levels of relevance" as structural elements of discourse evolved out of discussions between the writer and his thesis supervisor, and grew out of the intuitive notions of both. The literature provided remarkably little help in the general formulation of the theory. Much discussion was necessary before the elements of the theory were identified and suitably linked. From them were developed the necessary criteria for the description of text structure, as required by the thesis project.

For the scope and purposes of this thesis, "levels of relevance" will be virtually interchangeable with "levels of importance". These levels of importance develop as the result of the use of certain characteristic words, phrases, and constructions by writers or speakers, who in this way impose organization on their discourse (37). We, as readers or listeners, intuitively expect to find a certain number of sentences which carry and develop the main theme or train of thought of the author or speaker, and to find other sentences





whose function it is to outline and develop subthemes. These subthemes may be single phrases within a sentence, or a sentence itself, or a more complex group of sentences. Their function in the presentation might range from giving detail and example to fully developed explanations and descriptive passages. Therefore, each sentence we, the audience, encounter undergoes a value judgment to "place" it in relation to the main theme of the author's or speaker's communication. One might suggest that the extent of our "understanding" of that communication is proportional to our ability to replicate the structure from whence it came.

The last sentence implies that the author or speaker has consciously or unconsciously performed a synthesis analogous to the synthesis that he expects the audience to perform. If the communicator has written his speech or paper from a "sentence outline", he or she has performed a major part of the synthesis both consciously and formally. This concept of a sentence outline forms the basis for the description of the structure of discourse found in this thesis, and therefore will be discussed at some length in Section 2.2. For the purposes of the thesis, the theory of discourse will be developed primarily with reference to scientific and technical discourse, and the remainder of this subsection will offer further explanation for this restriction.

The point has already been made that it may be expected that information systems dealing with scientific





and technical documents, which are samples of scientific and technical discourse, will continue to be important. It has also been indicated that there will be some sort of public demand for the "translation" of scientific and technical discourse from one level to another for various purposes. In themselves these form valid reasons for confining this study to samples of scientific and technical writing.

There is, however, one further and very important justification for this restriction. The thesis proposes that style is defined by structure, developed from both form and content, and that styles of groups can be recognized through identities of structure. Exposition is more tightly structured than other forms, such as narration, and scientific writing is one of the most formal of the varieties of exposition. It is a type of discourse that lends itself readily to analysis and test of the basic hypothesis and the practical techniques developed in the thesis. It is very amenable to an analysis that is described through a sentence outline. Further, since writing "about" science and technology will run the gamut from the top level of the expository range to the bottom a series of samples of discourse may be tested, which will yet have common elements. In addition, from the practical point of view, the examination of this range of exposition should provide helpful insights for information systems concerned with documents dealing with other types of exposition.

It should be added that in imaginative literature these "levels of relevance" may be "played" back and forth to produce very special emotional effects. In a certain sense, then, the model and its explanation through the sentence outline may not be regarded as valid for all recorded discourse. However, it should be remembered that these emotional effects



are partially gained by a departure from the "normal" structure. At some later date it might be very interesting to apply the technique developed to special samples of creative writing.

## 2.2 The Sentence Outline

A sentence outline is an internal skeleton on which the body of a text is constructed. A hypothetical example of a sentence outline is given in Figure 3.

Composing from a sentence outline, the writer is limited in his choice of order to that described in the sentence outline. In the sample in Figure 3, the first paragraph (the paragraph is the usual unit assumed for the largest interval) contains the sentences A, A1, Ala, Alb, A2, and A3, and we would expect the ideas in them to occur in that order in the final form of the communication. Should the author, in some later draft of the paper, wish to lengthen the first paragraph, he could accomplish it in two basic ways. The first would be to add more sentences to existing levels. Thus the paragraph A, A1, Ala, Alb, A2, A3 might become A, A1, Ala, Alb, A2, A3, A3a, A3b, A4, A4a, or A, A1, Ala, Alb, A2, A2a, A2b, A3, . . . or some similar configuration which would conform to the general original pattern of the sentence outline. The second means of enclosing more information would be to add another level of "depth" to the structure in the sentence outline itself. The sample outline A, A1, Ala, Alb, A2, A3 could be expanded to A, A1, Ala, Alai, Alaii, Alb, A2, A3 should details concerning one of the disadvantages be included. The hypothetical sample might then occur as Figure 4.





- A. The process employed in the analysis was cryogenic fractional holography.
  - 1. This is not the typical industrial method because of two disadvantages.
    - a. First, it is too time-consuming because the test has an 18 month gestation period.
    - b. Second, at  $5 \times 10^6$  units per gram-liter, it is too expensive.
  - 2. This team used the method because Gamma University equipment was available.
  - 3. The experiment will be continued when the new Omega Technical Institute laboratories are installed.
- B. Polarized micro-sections of aurifous cellulose were chosen as the sample material.
  - 1. These organic compounds appear as by-products in  $\text{MnCO}_k$ -class reductions.
    - a. Commercial applications of this method could be economically advantageous.
  - 2. ....
    - a. ....
    - b. ....
- C. ....

Figure 3: A Hypothetical Sentence Outline



The deletion of sentences from a sentence outline also involves consideration of structure. In Figure 3, the sentence denoted 3. may be removed without disturbing the integrity or cohesiveness of the overall structure, but the same is not true of the sentence labelled 1. Should the author in a later draft wish to postpone discussion of the disadvantages of the process until a later paragraph, he must remove not only 1., but 1a and 1b, and, if present, 1ai and 1aii, and any other attendant structure.

### 2.3 Recognition of Structure

To make the value judgments that "place" each sentence into an overall structure, the reader or listener uses all the information available to him in the assessment. In oral communication this would include the speaker's emphasis, modulation, and if visible, gestures. In written communication, the information available about the inter-sentence relationships is concentrated in the sentence connectives, those words, phrases, and constructions which link together the sentences into connected discourse and help to inform the reader of the relative importance of each of the sentences. In classes on rhetoric some of these connectors are called "transitional" elements (38).

This ranking of the sentences of the text is not generally attempted en masse; it is extremely difficult to look at a particular sentence chosen from the beginning of a text and compare its relative import to a sentence selected



A. The process employed was cryogenic fractional holography.

1. This is not the standard method because of two disadvantages.

a. There is an 18 month gestation period.

i. This does not include planning & preparation.

ii. It is unnecessary to have skilled help until the 14th month.

b. It is frightfully expensive.

2. ....

.  
.  
.

Figure 4: Sentence Outline with Additions





from the nether end of the text. What is attempted with each incoming sentence is some kind of localized ranking with respect to the other sentences in some neighbourhood, and this analysis continues until the text is exhausted.

A more workable measurement than "relative importance", which signifies a value judgment, is the "relative dependence" of each sentence. Consider the three unrelated sentences of Figure 5. These sentences may be ranked from least dependent to most dependent, or most independent to least independent with respect to the textual environments from which they have been selected. They are also assumed to be ranked in order of decreasing importance to the author's main theme. The nature of the correspondence between the independence and the importance of sentences in text is demonstrated by Figure 5; intuitively we expect Sentence 1 to be more important to its author's presentation than Sentence 3 is to its author's presentation, although without the complete texts containing both and some value judgments it is impossible to be certain. In this thesis, the relative dependence of a sentence will be taken to be a reliable indicator of its relative importance.

The recognition procedure that gives the structure of the presented discourse will consist of an examination of the sentences for relative dependency, an examination of the sentences for intersentence links in terms of content or subject matter, and means of correlating this information. Each of these components will now be considered in detail.



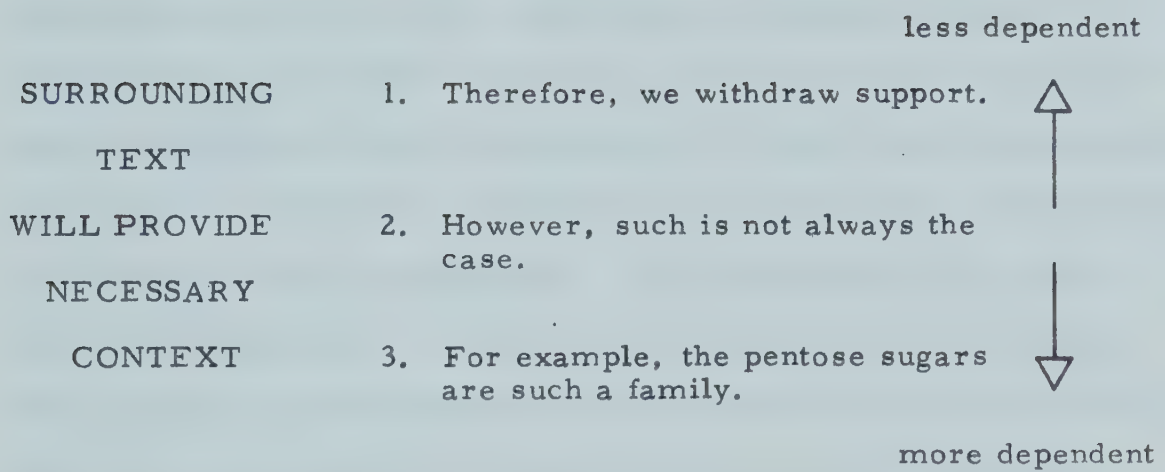


Figure 5: Sentence Dependency Hierarchy



### 2.3.1 Dependency

The establishment of the dependency hierarchy in Figure 5 is achieved through recognition of the connectives "Therefore", "However", and "For example". In general, each sentence in text can be examined for such words and phrases, which act as indicators to the relationships involved. There may be none or many of these indicators in each sentence. The impression given by each indicator in a sentence need not be the impression rendered by other indicators in the same sentence, and individual indicators may not always be used consistently. Since some indicators are judged to be more reliable than others, more significance is placed on their occurrence. It is proposed that the ultimate dependency of each sentence in text be correctly established by the summation of the occurrences of all the indicators, of whatever significance, in that sentence.

Three dependency categories are defined and recognized by the analysis. The first is typified by Sentence 1 of Figure 5, and sentences placed in this category will be basically independent sentences. These will be termed superordinate sentences, and, as previously indicated, will be presumed to be the most important in the author's presentation. In the sample sentence outline (Figure 3) the superordinate sentences would be A, B, C, . . . and all others on this primary level.

The next dependency category is typified by Sentence 2 of Figure 5. The sentences placed in this category, the





coordinate category, will be of equal dependence, and of equal presumed importance, with other sentences in their immediate environment. Examples of sentences in a coordinate relationship are Sentences 1a and 1b in Figure 3. Further examples of coordinate linkage would be Sentences A1, A2, and A3 in the same Figure 3.

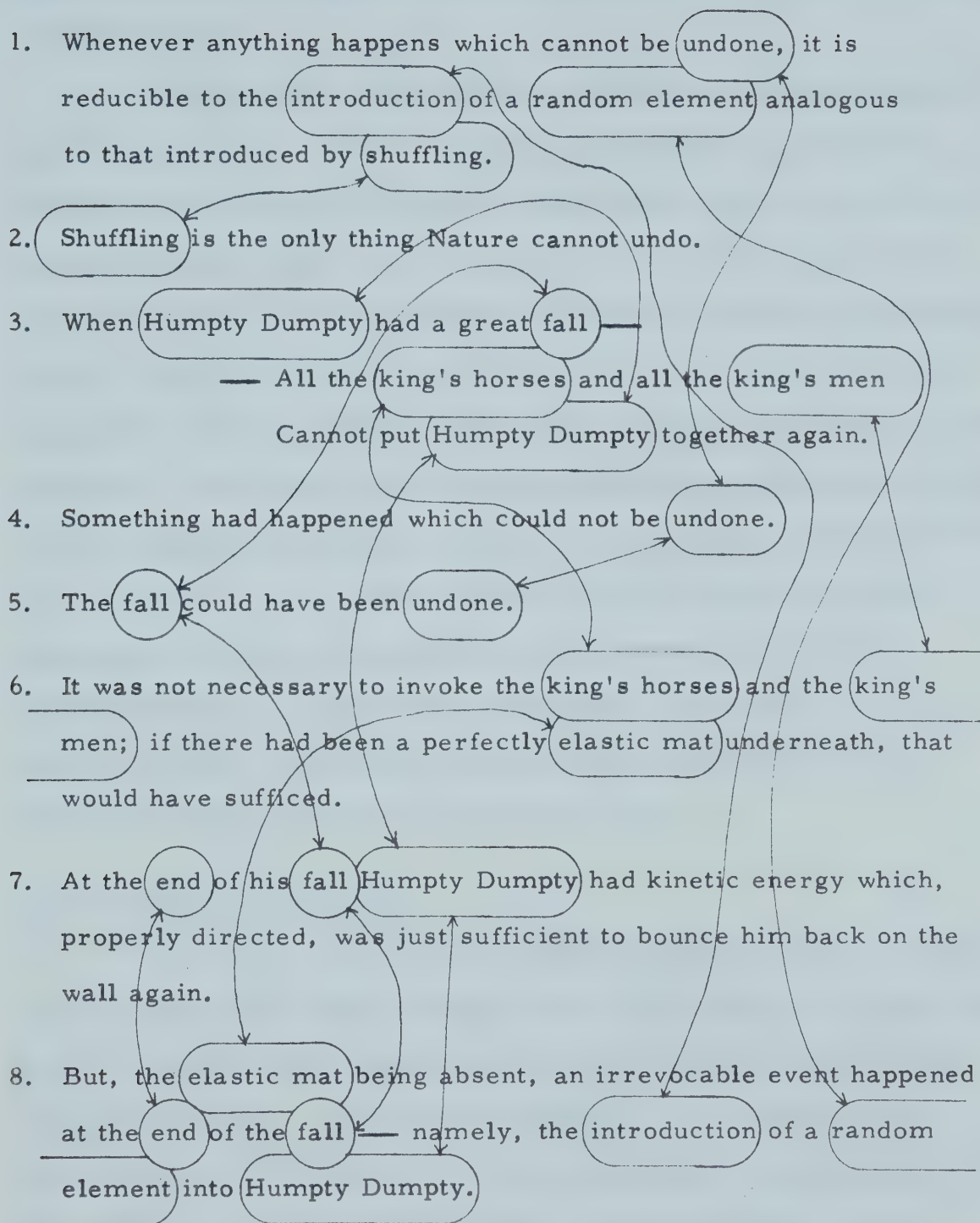
Sentences placed in the third dependency category are typified by Sentence 3 in Figure 3. Placed in this subordinate category are sentences which have been found to be dependent on other sentences. In Figure 3, subordinate links exist between A and A1, and A1 and A1a. Indeed, subordination is found whenever there is an increase in depth.

It is postulated that each sentence can be placed in one of these three relative dependency categories. The criteria for the final placement is the cumulated impression upon the receptor of certain indicator words and phrases. This is translated into the sum of the weighted occurrences of these indicator words and phrases.

### 2.3.2 Content Relations

It is postulated that structure is expressed not only through dependency but also through content links. Therefore, the second component of the recognition procedure is the examination of the environment or neighbourhood of each sentence, during which sentences closely related in terms of content are sought. The process is illustrated in Figure 6. In this example, however, the content matching is performed





from The Nature of the Physical World

by Sir Arthur Eddington

Figure 6: Content Linking



over the entire sample, rather than over a selected fixed neighbourhood.

In practice, the content linking is accomplished by matching all the "content words", as opposed to function words, with those contained in sentences within a predetermined distance from the sentence in question. Sentences containing one or more matches on content words are defined to be "content-related", and the more occurrences the two sentences have in common, the more strongly are they content-related. The size of the neighbourhood or environment chosen for the analysis helps to determine the depth and complexity of structure. A neighbourhood of radius equal to three sentences admits more possible configurations than a neighbourhood of radius two sentences (see Figure 6); a radius of four sentences involves more complexity than a radius of three sentences, and so forth.

### 2.3.3 Pattern Assembly

The first two steps in the procedure yield a collection of sentences each tagged with a dependency category and with a list of their most-related neighbours according to content linking. From this information is established a "most probable" category for each sentence, consisting of a type (Superordinate, Coordinate, or Subordinate) and a closest relative, based on content-linking. These most probable states might be regarded as the terminal symbols for the set of "production rules" which assemble the sentences





into a pattern. Figures 7a and 7b demonstrate the relationship between the finished pattern and the underlying sentence outline. Figure 7c is an alternate and interchangeable form of 7b. It is given to illustrate that the analytic and synthetic procedures could be used to produce varying forms of graphic output.

Figure 7b represents an eleven sentence text, the nodes being the sentences numbered in the order that they appear in the text. The top level (Sentences 1, 2, 7, 8, and 10) represent the main theme, or principal line of thought, in the sample text. The lines joining these nodes are coordinate links, indicating a relationship between sentences of equal relevance. The first substructure (Sentences 2, 3, 4, 5, and 6, contains two subordinate links, one between nodes 2 and 3, and the other between 4 and 5. In each instance, the sentence on the lower level has been established by the analysis as in some way subordinate to the content-related sentence on the upper level. The series 3, 4, and 6 represent what might be called a subtheme. They are connected by coordinate links, with equivalent importance attached to each. The second substructure (nodes 8, 9, and 11) may be interpreted similarly. The overall structure of the sample text suggests a paragraph split around Sentence 7, and that the last two sentences might profitably be rearranged, since Sentence 11 develops the subtheme of 9 and the text ends logically with Sentence 10.



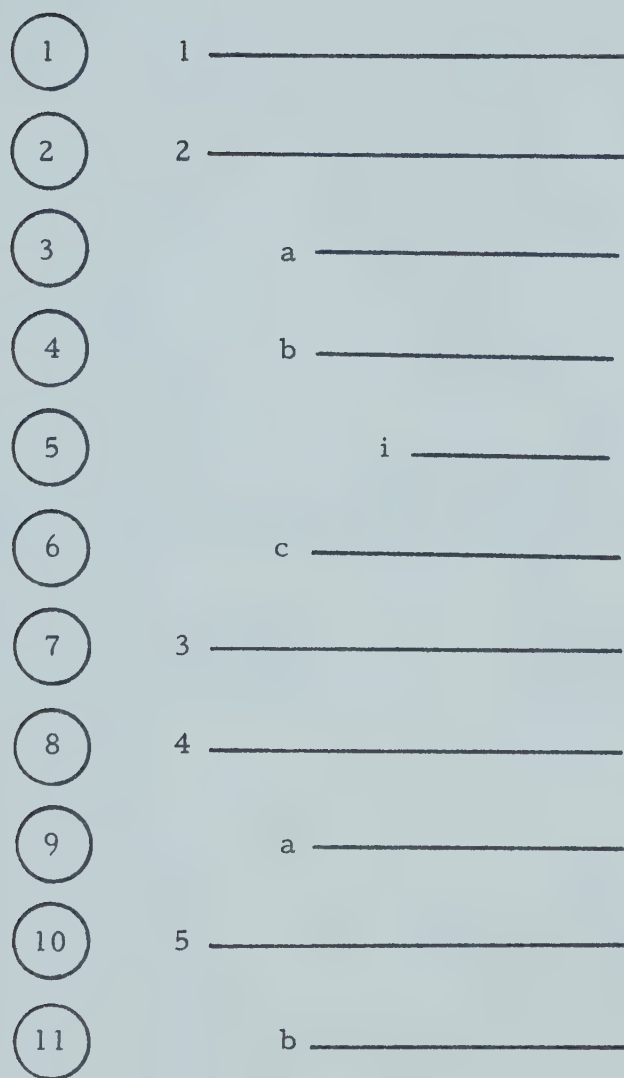


Figure 7a: Sentence Outline



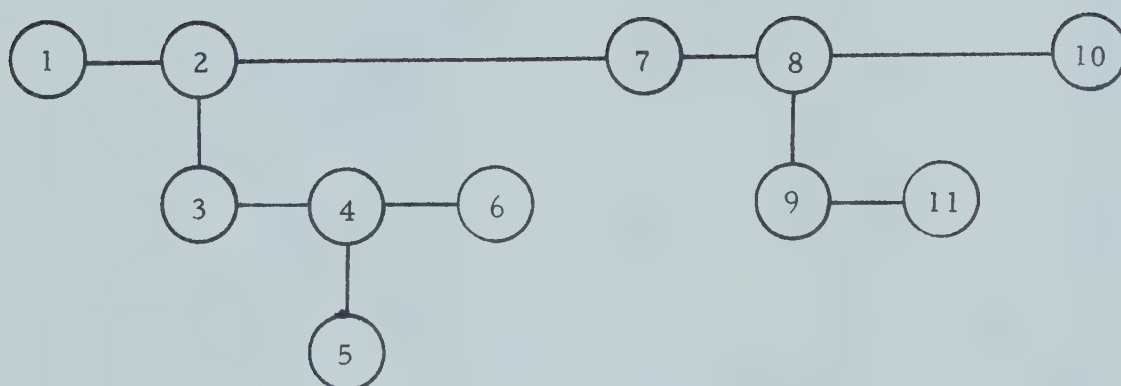


Figure 7b: Sentence, or Text, Diagram; Equivalent to 7a and 7c







Figure 7c: Sentence, or Text, Diagram; Equivalent to 7a and 7b



## CHAPTER III

### IMPLEMENTATION OF THE MODEL

#### 3.1 Introduction

The algorithm for the determination of text structure discussed in the previous section initially suggested two computer implementations, one employing unedited text and the other employing "augmented" text. Therefore, two procedures were defined and developed manually to a point where a small sample text could be analyzed according to a system of flowcharts; then the results using each method were compared. This step was important and, as it involved the search for suitable texts, it took considerable time. However, since it was essentially part of the necessary preliminary investigation it is not discussed in detail here.

The comparison of the manual analyses showed that the procedure employing "augmented" text was only marginally more effective than the procedure using unedited text, and to improve the system at all significantly would take complex and sophisticated techniques completely outside the bounds of the project as originally envisaged. On the other hand, the unedited version uses simple procedures; these are matching, maintenance of lists, tagging, and assemblage. The procedure using unedited text was, therefore, subsequently developed into the computer implementation PLATEXT; the procedure using augmented text was dropped. Both procedures



are described in the sections following.

The computer program associated with this thesis has been very thoroughly documented because it does not readily admit comparison with existing programs; in fact, it does not fit into any recognized niche as far as programming application classification goes. Its closest relatives are the larger examples of programs for content analysis but it is not a derivative of any or of all of them. The second reason for a detailed explanation is that it will hopefully make clear the constraints under which the program must operate. An example is the detail devoted to the sentence-termination-seeking and word-termination-seeking routines. These routines, as with all free-text handling programs, received a significant fraction of the programming effort, and even yet mishaps in these routines, along with array overflow, account for almost all program failures. A third reason for detailing the implementation to this extent is that the program might well find application outside the normal borders of computing science.

### 3.2 General Procedure for Unedited Text

The computer implementation of the unedited text version of the procedures developed was designed to utilize the computing facilities at the University of Alberta Computing Centre, which include a Model 770/663 CalComp plotter. The Centre's IBM 360/67 supports a number of languages under (initially) OS, and since May 1970, MTS.





The programmed model is called PLATEXT (PLotted Analysis of TEXT) and is written in FORTRAN IV augmented by a Computing Centre Library routine. The routine, ICMPAR, is an assembler string comparison package which enhances the text-handling capabilities of FORTRAN. PLATEXT also uses CalComp sub-routines and it was designed to do so from the outset of programming; the CalComp routines are most readily accessible from FORTRAN and this consideration determined the choice employed in the computer realization of the theory; programming details will be left to later sections.

The overall sequence of text operations is outlined in flowchart, Figure 8. The physical input to the program is a segment of text punched on cards using the standard EBCDIC character set. The text is unedited except for the deletion of headings, captions, numbering, etc. and can be keypunched onto the cards exactly as it appears in print. The text sample may include hyphenation, irregular spacing, indentations, and so forth. The limitations on word length, sentence length, and sample length will be discussed in the sections on programming details. Individual sentences must be separated by at least two blanks.

The text is read in and undergoes initial dependency and content relation analysis on a sentence-by-sentence basis, and, for the remainder of the analysis, is considered as a whole. The steps involved are, broadly:

- 1) Text is read into core until the end of the current sentence is reached.



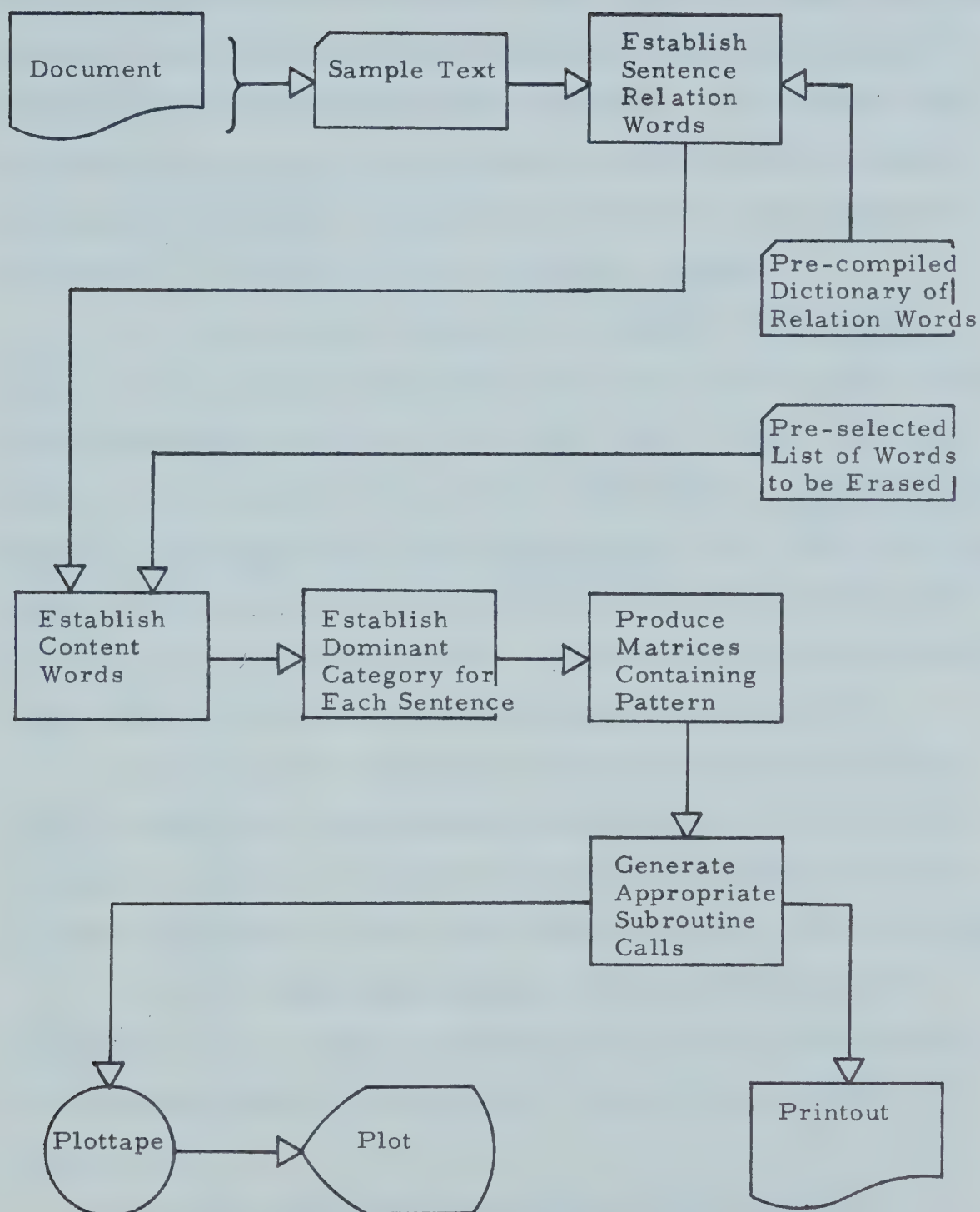


Figure 8: Sequence of Text Operations



2) The in-core sentence is scanned for words occurring in the dictionary of relation words. The dictionary is a pre-compiled list of dependency indicators, with a weighting or reliability factor associated with each. The contents and weighting factors of the dictionary were established empirically through manual analysis of scientific and technical literature. Details of any matches are stored.

3) The in-core sentence is examined for content words. These are the remainder after the words of the sentence have been passed against a deletion list of about 500 most-common words compiled from Kucera and Francis, A Computational Analysis of Present Day American English (39). The remaining words, content words, are matched against the content words of the two previous sentences. Details of any matches, plus the content words of the in-core sentence, are stored.

4) After all sentences of the text have undergone steps 1 through 3, the stored information on each sentence is used to select a "dominant category" for that sentence -- a coded estimation of the most likely role for that sentence.

5) The coded descriptions for each sentence are assembled into a pattern according to a set of simple rules for representation of structure. The pattern is coded in two matrices.

6) The plot production module translates the contents of the matrices into subroutine calls, which output the appropriate plot commands onto a plottape.

7) The pattern is produced on the CalComp plotter,





and a copy of the text indexed by sentence number plus some pertinent text statistics are printed on the line printer.

The details of the program PLATEXT are described in the following section.

### 3.3 Programming Description

#### 3.3.1 PLATEXT Main Program

3.3.1.1 BLOCK DATA. The BLOCK DATA segment of PLATEXT sets out those variables and arrays whose contents and size are known prior to the running of the program, and which will not be changed during execution. The allocation of storage and the initialization of values is done by the compiler, so that an object module will contain these present values.

BLOCK DATA contains:

AST	* bbb							
BLANK	bbbb							
FINAL	Z bbb							
ONE	1 bbb							
TWO	2 bbb							
THREE	3 bbb							
ALPHAN	A	b	b	b	B	b	b	b
	C	b	b	b	D	b	b	b
	E	b	b	b	F	b	b	b
	G	b	b	b	H	b	b	b
	I	b	b	b	J	b	b	b
	K	b	b	b	L	b	b	b
	M	b	b	b	N	b	b	b
	O	b	b	b	P	b	b	b
	Q	b	b	b	R	b	b	b
	S	b	b	b	T	b	b	b
	U	b	b	b	V	b	b	b
	W	b	b	b	X	b	b	b
	Y	b	b	b	Z	b	b	b
	0	b	b	b	1	b	b	b
	2	b	b	b	3	b	b	b
	4	b	b	b	5	b	b	b
	6	b	b	b	7	b	b	b
	8	b	b	b	9	b	b	b
	\$	b	b	b				



```

SPCHR      : bbb ; bbb ( bbb ) bbb " bbb ? bbb
            , bbb . bbb ' bbb bbbb - bbb ! bbb

```

BLOCK DATA also sets aside room for the erasure (or stop) list and the index to the erasure list (ERASUR(2320) and LINERA (480) respectively). The content of these is fixed and hence can be entered during compilation of PLATEXT. On the other hand, the dictionary of relation words has a relatively fluid content, and is entered during the execution phase.

3.3.1.2 Dimensioning and Initialization. PLATEXT has three classes of storage area: the first, (a), is the labelled COMMON areas set up by the BLOCK DATA division; the second, (b), is the COMMON areas defined by the main program; and the third, (c), is arrays and variables local to the main or subroutines and defined therein.

(a) These were discussed in the previous section.

(b) The MAIN segment of PLATEXT defines in the COMMON area two arrays HORZ and VERT, which serve as an interface between the analytic and synthetic functions of the program. The arrays are declared INTEGER and are dimensioned as 10 x 40 and 40 x 10 respectively. This anomalous dimensioning is a reflection of their function in the program: HORZ will contain as elements sentence numbers which are to be linked horizontally, so that on printout it is helpful to arrange them by row (i.e. xxxxxxxx ); VERT as the name implies contains as elements sentence numbers which are to be linked vertically, and the least confusing way to



x x  
x x

display them is by column (i.e. x x ).

x x

The dimensions (10 and 40) were established after a number of trials that employed smaller arrays produced truncated patterns. Despite the present limit on the number of sentences in each sample, it is not inconceivable that an extreme sample might overflow one boundary or another if the majority of its sentences are connected by the same kind of link. The cheapest insurance against this is the provision of large arrays, with the acceptance of the fact that they will be but sparsely filled. An alternative is some form of linked-list structure but the amount of referencing done to HORZ and VERT suggests that the process of addressing elements should be kept as simple as possible. HORZ and VERT are initialized to zero.

(c) The MAIN segment of PLATEXT defines 19 arrays which do not appear elsewhere in PLATEXT. A brief description of each follows:

1. DATA (1024). This is the plotter output buffer referenced in the opening statement CALL PLOTS (DATA (1), 4096). (See subroutine PLOTS.) The array is not initialized.

2. DICTN (80,25). This array will contain the dictionary of relation indicators. With the quoted dimensions the dictionary will accommodate 80 entries of up to 22 characters each (the remaining three characters being a numeric code). At the present writing the longest dictionary entry is 17 characters (ON THE OTHER HAND); an addition to the dictionary of up to 5 characters longer than this example





can be accommodated without redimensioning the dictionary. The dictionary array is not initialized.

3. CARD(160). This array functions as an input buffer for the text, presuming that the text is being read in from cards, and is the maximum amount of text that has to be stored in core at once, excepting the content word buffers (see below). The initial search for a sentence boundary occurs in CARD, and the test for the end of the input text will also be performed in this buffer. The area is not initialized.

4. SENT(500). This array is the principal work area of the program. After a sentence boundary has been established in the input buffer, the sentence will be moved, in fragments if necessary, to SENT for analysis. SENT is not initially blanked, but, as through the course of execution various amounts of SENT are filled from the input buffer, unused portions of SENT will be blanked. This tidying-up is unnecessary for the analytic procedures but allows SENT to function as an output buffer for the sentence without hindering its role as the residence of the sentence while its constituent words are found and analyzed.

The dimension of SENT(500 characters) represents the upper bound, plus a bit, of sentence length so far encountered, save one example. As the single incorrigible sentence ran to 300 words and over 1600 characters, one might safely pronounce it divorced from the mainstream of scientific and technical writing, and henceforward ignore its existence.



5. SLIST 1(50,25), SLIST 2(50,25), SLIST 3(50,25), ISNTWD(3). These arrays are the three content word buffers and their associated index. Each can accommodate up to 50 words of up to 25 characters each, a limit which at least to this date has proved more than generous; the maximum number of content words found within a single sentence so far is 32. The number of content words currently residing in each buffer is stored in an appropriate location of ISNTWD.

The content word buffers are initially blanked and ISNTWD is initially zero-filled.

6. WORD(25). This is the area into which words from SENT are assembled, one character at a time. The limit on word length imposed by the current dimensioning has been the cause of more concern than boundary overflow has warranted. Not even the, admittedly few, samples of scientific and technical literature dealing principally with organic chemistry have incurred problems with excessive word length, although the maximum was approached more closely in this subject field than any other yet encountered. There are not many words in English that are more than 25 letters long.

WORD serves as the vehicle for initial matching on dictionary and content word buffer entries. It is not initialized.

7. IBLEEP(50). This array contains the number of words, not only content words, found in each sentence. The tallies are indexed by sentence number and incremented as each word is isolated. The array IBLEEP does not enter into the



analytic procedures of PLATEXT but is basic to the statistical segment of the program. IBLEEP is initialized to zero.

8. DUMMYD(25). This is a work area used for holding entries from the dictionary or content word buffers while comparing said entry to the current occupant of WORD. (See library subroutine ICMPAR.)

9. IXRT1(150,2), IXRT2(150,2), IXRTC1(50), IXRTC2(50). These arrays are the scratchpads and scratch-pad indicators on which the dictionary matches and content links are first recorded. IXRT1 contains as elements up to 150 pairs of values, one for each discovered dictionary match; specifically, IXRT1(B,1) indicates whether the Bth dictionary match was subordinate, coordinate, or superordinate, and IXRT1(B,2) contains the weighting factor associated with the dictionary entry involved with the Bth match. The indexing array associated with IXRT1 is IXRTC1(N), in which the Nth element is the number of dictionary matches (each detailed in IXRT1) found in the Nth sentence. Both IXRT1 and IXRTC1 are initialized to zero.

IXRT2 details the result of the search for matches on content word. If Sentence 24 contains the content word WXYZ and WXYZ reoccurs in Sentence 26, the latter occurrence (say the Qth occurrence of a content word in the text) will cause the following entry: IXRT2(Q,1) = 24 and IXRT2(Q,2) = 1. If WXYZ occurred twice or thrice in Sentence 24, IXRT2(Q,2) would have been set to 2 or 3; if WXYZ occurred twice in Sentence 26, there would have been





two separate entries IXRT2(Q,1 & 2) and (say) IXRT2(P,1 & 2) where each of IXRT2(P or Q,2) would reflect the number of occurrences of content word WXYZ in Sentence 24. As before, IXRTC2 is an indexing array where IXRTC2(N) is the number of paired entries in IXRT2 associated with the Nth sentence. Both IXRT2 and IXRTC2 are initialized to zero.

10. ICARD(20). This is an input area used when reading under an I4 format, i.e. 20 values per card. (See library subroutine ICMPAR.)

11. OVRL(50,2). This array contains the summarization of the dictionary and content word searches. OVRL(K,1) and OVRL(K,2) represent the conclusions re the Kth sentence: the former (1 = subordinate, 2 = co-ordinate, 3 = super-ordinate, 4 = no dictionary matches) is the result of the dictionary search for relation-indicating words, and the latter (1 = content link to previous sentence, 2 = content link to sentence twice previous, 3 = no content links found) is the result of the buffer search for matches on content words. OVRL is initially zeroed.

12. INDX(50,3). This array serves as an inverted index to HORZ and VERT. (See "COMMON areas defined by MAIN".) These two arrays, while of paramount importance as the interface between analytic and synthetic parts of the program, are not designed to be referenced by sentence number, having been constructed to answer questions of the form "Which sentence symbol is in this position?" rather than "In which position is this sentence symbol?".



INDX provides the needed facility for such referencing.  $\text{INDX}(J,1)$  indicates whether the  $J$ th sentence is to be found in HORZ ( $\text{INDX}(J, 1) = 1$ ) or VERT ( $\text{INDX}(J,1) = 2$ ) while  $\text{INDX}(J,2)$  and  $\text{INDX}(J,3)$  give the coordinates in HORZ or VERT of the  $J$ th sentence. INDX is initially zero-filled.

13. JLYST(50). This array is created and referenced in the segment of the program which reads from HORZ and VERT the sentence symbols to be plotted. JLYST(V) contains the sentence number of the  $V$ th sentence symbol plotted, and the array is initially zeroed.

3.3.1.3 Input Files. PLATEXT requires at this writing three files to be entered into the system: the erasure or stop list, the dictionary of relation-indicating words, and the text to be analyzed. Currently all of these are entered through the card reader.

a. The erasure list consists of a linear file of 2360 characters and an associated address index. Provision has been made for the specification of the list in the BLOCK DATA division; a dissatisfaction with the contents of the list means that the list at present exists in a temporary form (see discussion in Chapter VI) and instead of being resident in BLOCK DATA is entered at each run. Given a demonstratably suitable frequency count from which to draw the erasure list and a more defined sample universe, it would be profitable to enter the erasure list in the BLOCK DATA division.



The format under which the erasure list is entered will be discussed in the section on library subroutine ICMPAR.

The contents of the erasure list, with comments, form Appendix B.

b. The dictionary of relation indicators is kept on cards, one entry per card. This arrangement has proved very satisfactory in view of the fact that much of the "tinkering" with the program has been with the dictionary entries and with the weighting factor associated with each of them. The time spent to read in the 60 to 70 cards each program submission is a reasonable trade-off for the convenience of a readily adjustable dictionary file.

The format under which the dictionary file is entered will be discussed in the section on library subroutine ICMPAR.

The contents of the dictionary of relation indicators, with comments, form Appendix A.

c. The input text is entered as required during the analysis phase of PLATEXT. The initial entering of text is 160 characters, to fill the array CARD. As sentences are found in the search through CARD and transferred to SENT, a count is kept of the progress through the input buffer; when another 80 characters of text can be accommodated, the latter half of CARD becomes the first 80 characters, and the last 80 characters of CARD are filled from the input file.







The entering of the sample text through the card reader is appropriate in this pilot situation, but one would expect magnetic tape to be the typical input device for this application in the future, especially as more and more data bases become available in this medium. The modifications to the program necessary to adapt it for use with magnetic tape would include the provision of a larger input buffer to minimize I/O events.

The sequence of tests used to establish sentence boundaries will be discussed in a later section.

The entering and searching of text is terminated and the next phase of the analysis begun when an end-of-text symbol "\*\*\*" is encountered.

3.3.1.4 Sentence Boundary Definition. The text entered into the buffer CARD is examined sequentially character by character. All characters are initially checked for the end-of-text symbol "\*\*\*", and subsequently investigated for the possibility of a sentence-ending sequence. The procedure can be grossly described as selecting those "special characters" from SPCHR which could be involved in a sentence termination, and exhausting the possible occurrences. (See flowchart, Figure 9.) The vast majority of sentence endings fall into the "period, plus at least two blanks" category, and if one modifies that to read "full stop, i.e. period, question mark, exclamation mark, plus at least two blanks" the sentences not covered are very infrequent. The sentences remaining are most often examples



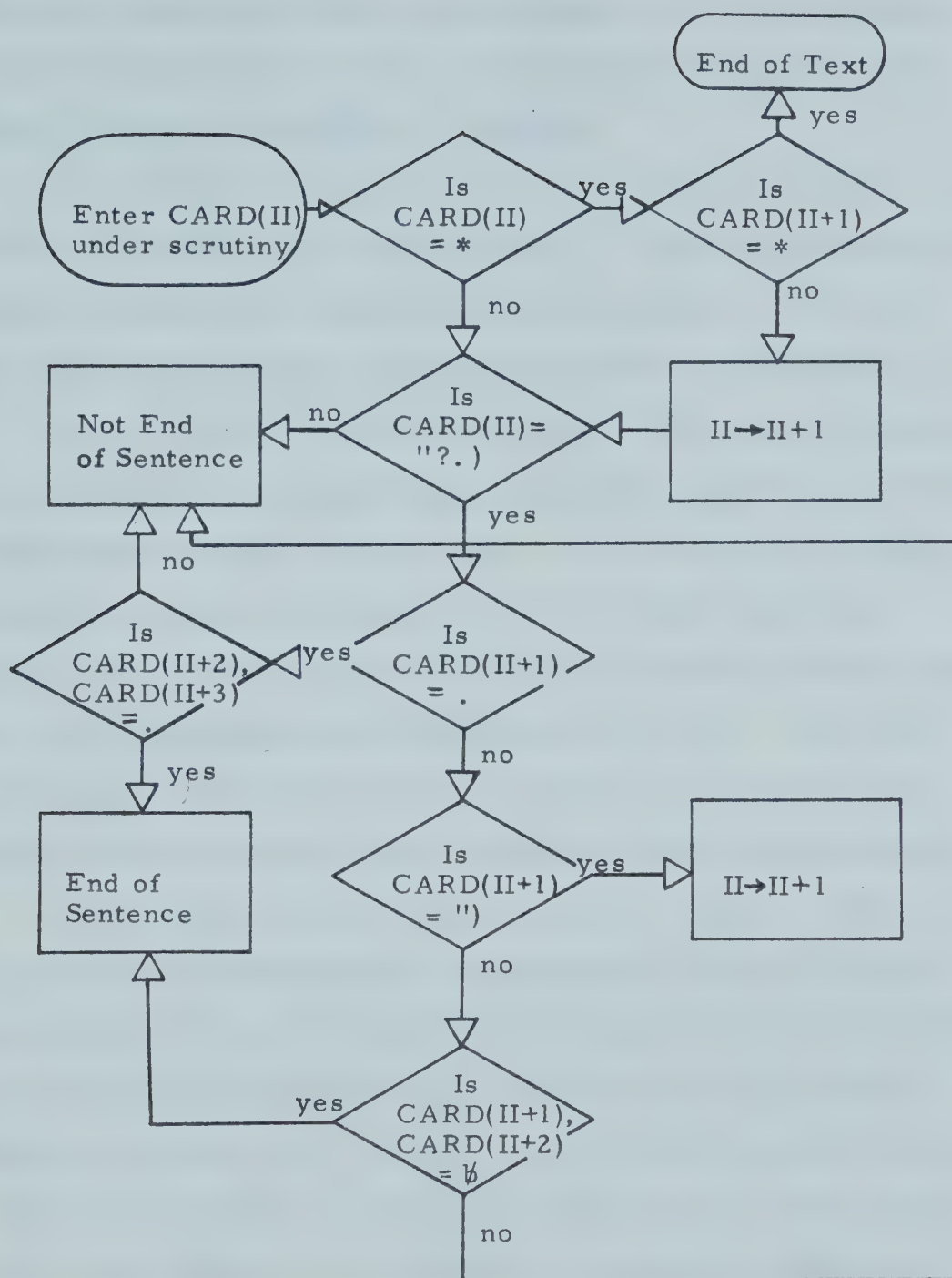


Figure 9: Sentence Isolation Routine



containing quotation marks, parentheses, or a combination of the aforementioned. This end-of-sentence testing is typical of any text-handling routine.

The latest sentence-terminating device to be included in PLATEXT was the ellipsis. Given the richness of English in regard to conceivable punctuation, it is not impossible that further modification might be required.

The sentence boundary discovery situation at present is not flawless. Abbreviations can still, under particular circumstances, cause "false drop". The limerick in Figure 10a would be found by PLATEXT to be of three sentences duration since the periods and spacing satisfy the requirements for a termination-of-sentence situation. No false drop occurs if the limerick is entered as in Figure 10b because the spacing no longer disguises the abbreviations.

The system "noise" invoked by the text of Figure 10a could be filtered using an exhaustive list of English abbreviations to check suspected sentence-terminating words. The two difficulties with that solution are also demonstrated by Figure 10a: firstly, no one can possibly foresee what will become an accepted abbreviation (five years ago Ms. was "message" or "missive"); secondly, sometimes sentences do end with an abbreviation.

The only method likely to yield a significant gain in performance in the situations described above would include some form of semantic analysis. Aspects of this problem are discussed in Section 6.3 of the Concluding





A GIRL WHO WEIGHED MANY AN OZ.

USED LANGUAGE I DARE NOT PRONOS.

FOR A FELLOW UNKIND

PULLED HER CHAIR OUT BEHIND

JUST TO SEE ( SO HE SAID ) IF SHE'D BOZ.

P. L. Mannock

Figure 10a: Sample Input I

A GIRL WHO WEIGHED MANY AN OZ. USED LANGUAGE I

DARE NOT PRONOS. — FOR, A FELLOW UNKIND PULLED

HER CHAIR OUT BEHIND JUST TO SEE ( SO HE SAID ) IF

SHE'D BOZ.

Figure 10b: Sample Input II



## Discussion.

3.3.1.5 Word Boundary Definition. After a sentence has been isolated and stored in the work area SENT, PLATEXT attempts to isolate the constituent words of the sentence. Many of the problems that occur in sentence definition have counterparts in word boundary definition. (Consider the line "WE LISTED FOR CONCORDING EVERY WORD IN DOBBIE'S TEXT EXCEPT THE FRAGMENTARY LE AT LINE 240 (KLABER'S (HWI)LE)."). It is perhaps not surprising that the word isolation process described by flowchart, Figure 11, should bear a family resemblance to the sentence isolation process described by flowchart, Figure 9.

Flowchart, Figure 11, describes the word recognition process in reasonable detail; the rest of this section will discuss one typical "policy" problem engendered by the attempt at word boundary definition.

A minor problem, caused by parentheses in text, provided the model for the solution to a major problem, caused by hyphens. Consider Figures 12a and 12b. It seems reasonable to have 12a consist of four words and 12b to consist of one word -- PLATEXT concurs.

The Figures 13 a-e indicate a more commonly-encountered difficulty: the hyphen and/or short dash. Each instance in Figure 13 illustrates a slightly different use of this overworked little symbol. Rather than try to legislate for each of these cases, the following guidelines were adopted. If the hyphen was encountered without an



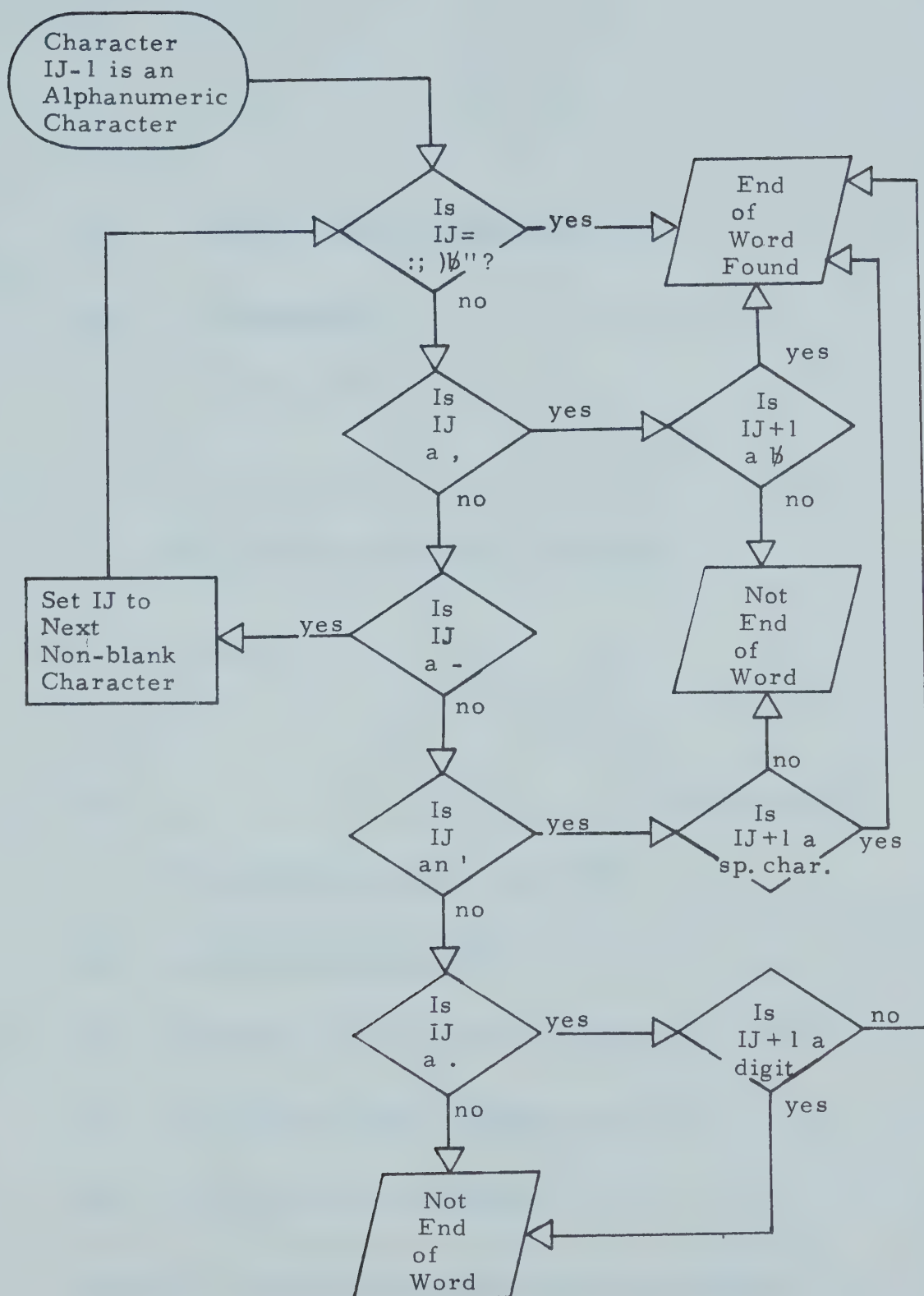


Figure 11: Word Isolation Routine





12a HENRY THE FORGETFUL (IV)

12b BWITCH(50)

### Figure 12: Parentheses; Two Uses

13a SUPERCALI-

FRAGILISTICXPIALIDOCIOUS

13b I. MACDONALD-SMITH

13c     $\text{PROFIT} = \text{REVENUE} - \text{EXPENSES}$

13d INCOME IS DOWN - TAXES ARE UP.

13e AND LEFT ME HANGING---

### Figure 13: Hyphen and/or Short Dash; Five Uses



intervening blank, the situation was assumed to be of the 13a type above. The hyphen and all following blanks are disregarded, so that the analysis restarts at the first non-blank character after the hyphen. As a result, examples 13a and 13b cause the following to be stored in the work area WORD:

13a        SUPERCALIFRAGILISTICEXPIALIDOCIOUS

13b        MACDONALDSMITH

The former will not fit in its entirety. For 13a, this is exactly the desired result. For 13b the result is not wholly satisfactory, but the most common instances of this are the "dapple-dawn-drawn" variety, and a condensation to "DAPPLEDAWNDRAWN" does not greatly constrict words already closely connected. The other alleviating argument is that the procedure is consistent so that "MacDonald-Smith" will always be matched against "MacDonald-Smith", irrespective of how the name ultimately appears in WORD.

If the hyphen is separated from the last non-blank character by one or more blanks, the hyphen is considered a punctuation mark rather than a connective, and is ignored. The words recorded out of 13c, d, and e would be:

13c        PROFIT REVENUE EXPENSES

13d        INCOME IS DOWN TAXES ARE UP

13e        AND LEFT ME HANGING

The drawback with this sytem is that the space preceding the hyphen is critical. If it is left out of 13c, the word "REVENUE EXPENSES" is stored. But 13e will still



sort itself out; HANGING---.→ HANGING.

As with the sentence isolation algorithm, modifications to the word isolation routine will doubtlessly suggest themselves as more samples are tested.

3.3.1.6 Matching of Dictionary Entries and Content Words. After each word has been isolated as described in the previous section, PLATEXT attempts to match the word against the contents of its two stored lists -- the dictionary file and the erasure list. If the word is not replicated within the bounds of the two lists, further matching is done against the files of content words from the previous two sentences. The process is outlined in flowcharts, Figures 14 and 15.

Most of the actual comparisons of these character strings is done by a library subroutine (ICMPAR) whose particular demands account for the extensive book-keeping of word-lengths in the program. Whenever in flowcharts, Figures 14 and 15, the comparisons between character strings have been achieved with ICMPAR, the fact is duly noted. ICMPAR and library subroutines in general are dealt with in separate sections.

The first step in the dictionary search process is the transferral of the first 25-character dictionary entry to the storage area DUMMYD. A comparison is then made between the n-character contents of WORD and the first n characters of DUMMYD. If, at any comparison, the contents of DUMMYD are found to "exceed" the contents of WORD, the





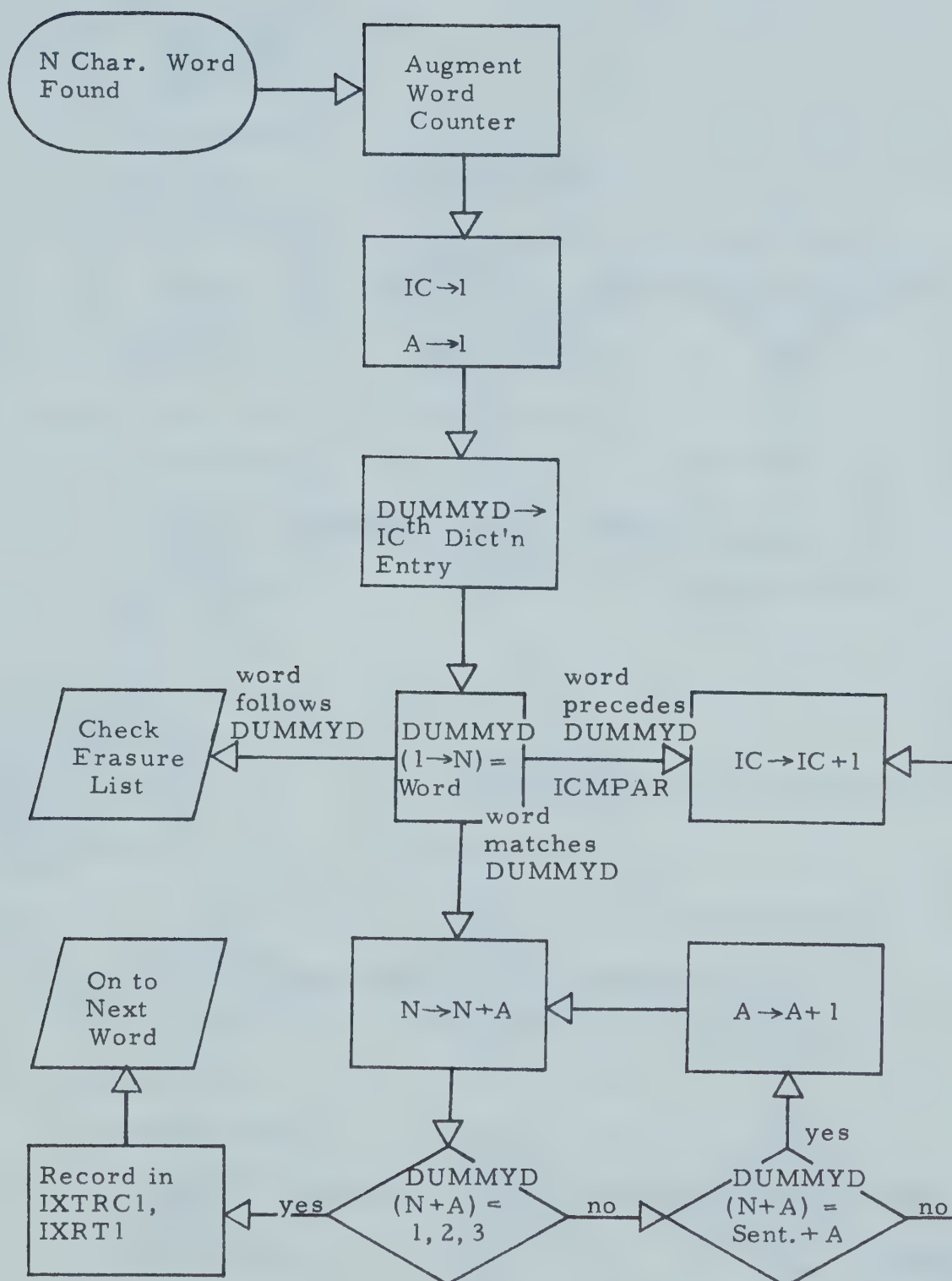


Figure 14: Dictionary Matching Routine







search is halted because the dictionary file is in alphabetical order and the contents of WORD obviously have no match therein. Should the opposite occur, the comparison moves to the next dictionary entry and ICMPAR is recalled.

If ICMPAR returns the result that the n-character contents of WORD and the first n characters of DUMMYD are identical PLATEXT examines the (n+1)st character of the dictionary entry in DUMMYD. If that character is a "1", "2", or "3" then the dictionary entry has been matched along its full length and the result may be recorded. If the (n+1)st character in the dictionary entry is not a numeric code, it and subsequent characters until the termination of the dictionary entry are matched directly against the first character in SENT after the n-character word. This character-by-character matching between DUMMYD and SENT is continued until either the dictionary entry is exhausted, or a difference is discovered. In the former, a match has occurred; in the latter, no match has occurred. If the latter event transpires, the next dictionary entry is read into DUMMYD and the process begins again.

Consider a sentence beginning with the phrase "ON THE OTHER SIDE...". The storage area WORD would initially contain "ON" and the word length indicator IWRDLN would indicate a 2-character word. This configuration would be tested against the dictionary entries, one of which would eventually be "ON THE OTHER HAND122". This particular dictionary entry would be read into DUMMYD and the first two





characters ("O" and "N") compared via ICMPAR to the contents of WORD. This will yield a match in the example under examination, so the next character in the dictionary entry is scrutinized. It is not a number, which would indicate the end of the entry had been reached, so the character ("blank") is compared directly against the character following "ON" in the original array SENT. Because this and subsequent comparisons are uniformly single-character, ICMPAR is not required. In this example, the third character will find a match and so the next character in the dictionary entry is examined. Being a "T" and not a numeral, the non-digit is compared to the next character in SENT. In this fashion the dictionary entry is matched from "ON" to "ON THE OTHER". The comparison of the dictionary entry's "H" to the array SENT's "S" will halt the sequence and analysis will revert to the next dictionary entry or the erasure list, depending on the number of dictionary entries beginning with "ON". The procedure is given in flowchart, Figure 14.

The content-word matching procedure is a two-step process and is outlined in flowchart, Figure 15. The first step is the passing of the contents of WORD against the erasure list. The search is sequential and is accomplished by using a table of relative addresses, i.e. word lengths, in conjunction with the linear erasure array. The 480 entries in the list are organized in order of descending frequency rather than alphabetic order, and, while the placing of the most probable words at the beginning of the erasure list



affords a meagre amount of economy, a binary search over a more structured list would save a great deal more.

If the word in question is found to be among those resident in the erasure list, PLATEXT reverts to word-seeking in SENT, with the pointers set to the character after the last character in the word just discarded, irrespective of any comparisons that may have been made by the dictionary matching mechanism on this and subsequent characters. In the example discussed earlier in this section, the phrase "ON THE OTHER SIDE" would be restarted at "ON, THE OTHER SIDE" after the investigation of the word "ON", even though that investigation required that the entire phrase be examined by the dictionary matching process.

If the word in question has not been encountered among the 480 words of the erasure list, it ranks as a "content word" and is stored in SLIST3.

The second part of the process is the matching of the newly-found content word against the content words of the previous sentences. In its present configuration PLATEXT matches over a radius of two sentences, so SLIST2 and SLIST1 contain the content words of the previous and twice previous sentences respectively. Each of the words in SLIST2 and SLIST1 is matched against each of the newcomers in SLIST3. The string-matching routine employed is again ICMPAR since the content words are of varying lengths.

A match is recorded for each occurrence of the content word in question: if a content word occurs once in





Sentence 1 and once in Sentence 2, one match is recorded; if a content word occurs once in one of Sentence 1 and Sentence 2 and twice in the other, two matches are recorded; if a content word occurs twice in each of Sentence 1 and Sentence 2, 4 matches are recorded.

The matches are duly noted in IXRTC2 and IXRT2, along with the earlier-discussed dictionary matches in IXRTC1 and IXRT1. When the sentence currently under scrutiny is exhausted of words, the SLIST2 array becomes the SLIST1 array, the SLIST3 array becomes the SLIST2 array, and finally SLIST3 is cleared for the content words of the next sentence.

3.3.1.7 Dominant Attributes. After all the dictionary-listed relation word clues and contentual links have been logged for each of the sentences in the sample, PLATEXT establishes a "most-likely state" for each sentence based on the contents of the arrays IXRTC1, IXRT1, IXRTC2, and IXRT2. These clues are usually many per sentence, and the algorithm employed to reduce the evidence is shown as flowchart, Figure 16.

Each sentence is placed in one of the twelve categories established by the 4 categories of sentence relative dependency (SUPERORDINATE, COORDINATE, SUBORDINATE, and NO INDICATION) furnished by dictionary matching, and the 3 categories of content relation (RELATED TO THE PREVIOUS SENTENCE, RELATED TO THE SENTENCE TWICE PREVIOUS, and NO INDICATION) furnished by matching within the content word buffers. The sentence relative dependency is established





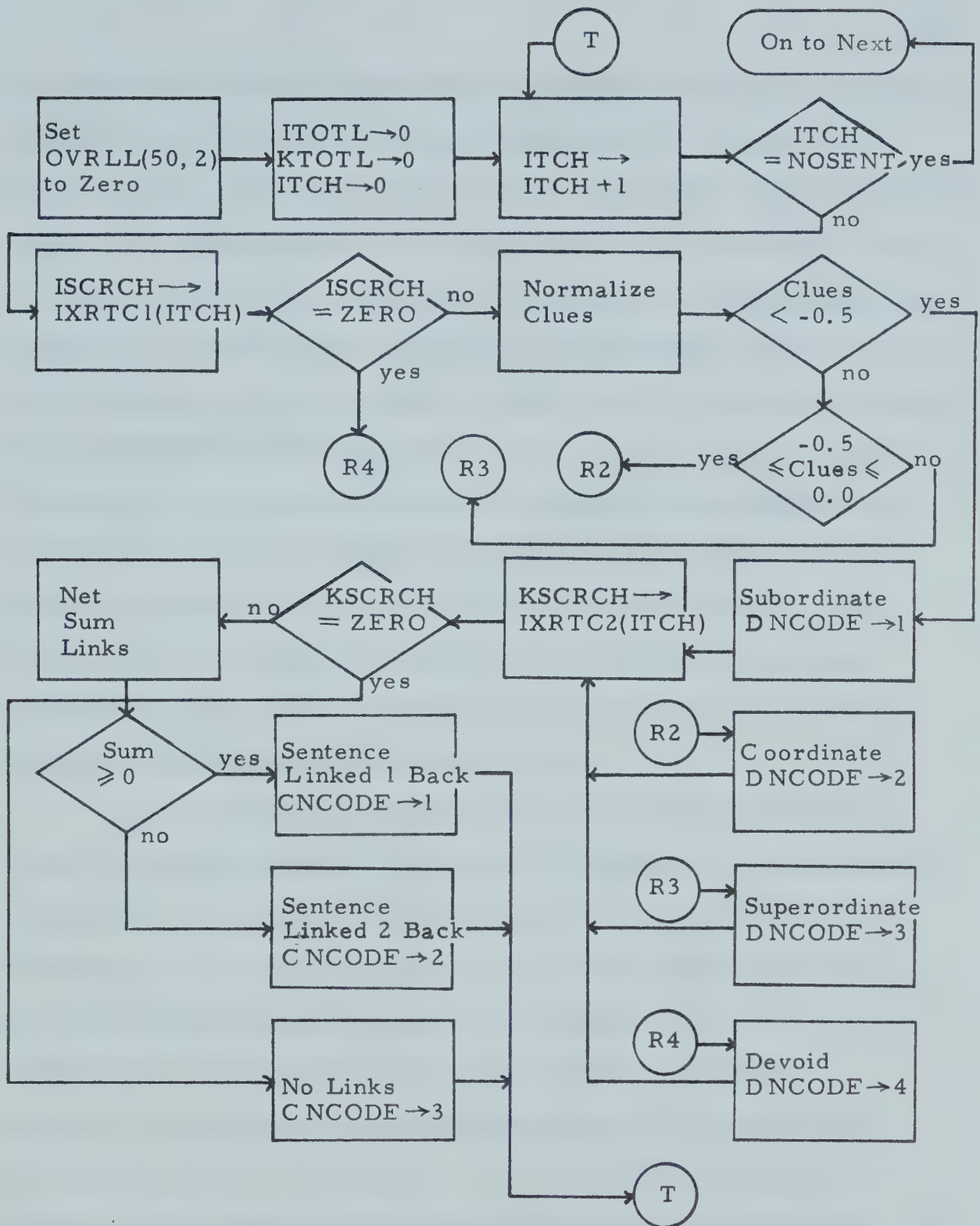


Figure 16: Establishing the Dominant Attributes of the Sentence



by adding the Clues times Weights for each sentence and then "normalizing" the results by dividing by the number of contributing clues occurring in each sentence. The normalized result for each sentence is projected on an adjustable scale: a figure around zero indicates a COORDINATE categorization, a figure well below zero indicates a SUBORDINATE categorization, and a strongly positive result indicates the sentence belongs in the SUPERORDINATE category. The division points of the scale may be adjusted, although in practice adjustments to the overall ratio of SUBORDINATE/COORDINATE/SUPERORDINATE have been unnecessary, most of the tinkering having taken place with the weights assigned to individual dictionary entries. If no clues occurred in the sentence, the fourth category (NO INDICATION) is selected.

The contentual relation of the sentence is established by totalling and comparing the number of matches with each of the previous two sentences. The sentence is declared to be content-linked to the antecedent sentence with the majority of matches. In the event the two antecedent sentences have an equal number of content matches the sentence immediately previous is selected as the content-related sentence. If no content matches in either of the two previous sentences have been recorded, the third category (NO INDICATION) is selected.

The dominant categorization of each sentence is stored in the array OVRL.

#### 3.3.1.8 Pattern Generation. The twelve



categories of dominant sentence characteristics provide the basis for five different kinds of sentence linking. The relationship is shown in Figure 17. A brief comment on each of the five follows:

1) Superordinate: When it has been established from dictionary matches the sentence under scrutiny is superordinate, i.e. less dependent, more relevant or important, the sentence is immediately placed on the top level and connected with the previous top-level sentence, whether the previous top-level sentence was the sentence physically previous, physically twice-previous, or neither of these. The action of a sentence heavily loaded with THIS, THEREFORE, etc. is to restore the sentence pattern to the principal line of thought or argument, irrespective of the level of dependency of the current or antecedent sentence.

An identical result is performed by a sentence containing no dependency indicators and no content word matches. The rationale is that a sentence with no linking of any kind to the previous sentence or sentences is very likely the introduction of a new topic or theme.

2) Coordinate with Previous Sentence, 3) Coordinate with Sentence Twice Previous: If co-ordinate links are dominant in a sentence, the sentence will be coordinate with either the previous sentence or the sentence twice previous, depending on the content link component. If there should be no indication of content linkage, the sentence is adjudged to be coordinate with the previous sentence, which





Content Relation:			
Sentence Relative Dependency:	Previous Sentence	Sentence Twice Previous	No Indication
Superordinate	SUPER.	SUPER.	SUPER.
Coordinate	CO. WITH PREVIOUS	CO. WITH 2ND PREVIOUS	CO. WITH PREVIOUS
Subordinate	SUB. TO PREVIOUS	SUB. TO 2ND PREVIOUS	SUB. TO PREVIOUS
No Indication	CO. WITH PREVIOUS	CO. WITH 2ND PREVIOUS	SUPER.

Figure 17: Table of Sentence Characteristics



is the most likely candidate for the encountered coordinate linkage.

If content links with one or the other of the two previous sentences indicate a strong link in terms of subject matter, but there is no indication of relative dependency, the sentence is presumed to be a continuation of the subject thread of the content-related sentence. A coordinate linkage is the appropriate vehicle for the representation of "continuation of subject", and so is the default choice when only content links are present.

4) Subordinate to Previous Sentence 5) Subordinate to Sentence Twice Previous: The two subordinate linkages are the straight-forward result of analyses that indicate subordinate relation words are the dominant feature of those particular sentences. As before, a lack of content links causes PLATEXT to default to "related to previous sentence"; in this case, "subordinate to previous sentence".

The four categories emerging from the dictionary comparisons (abbreviated to "super", "co", "sub", and "nil") and the three categories emerging from the content word comparisons (denoted "back 1", "back 2", and "nil") are pieced together by the procedure described by Figures 18 (two parts).

The necessity for a procedure as complex as that taken in the flowcharts, as opposed to a simple table look-up employing Figure 17, can be indicated through a hypothetical example.



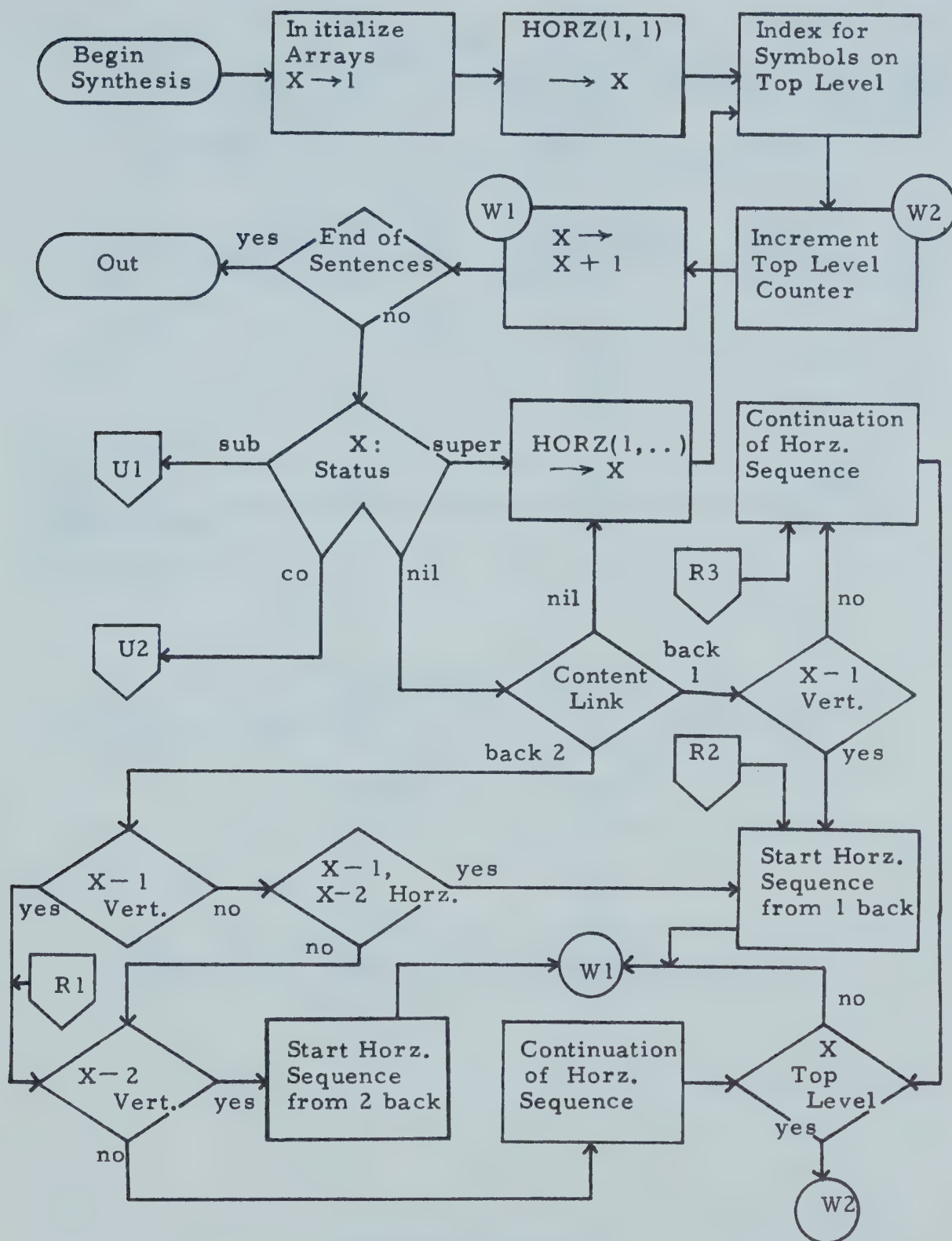


Figure 18: Pattern Synthesis





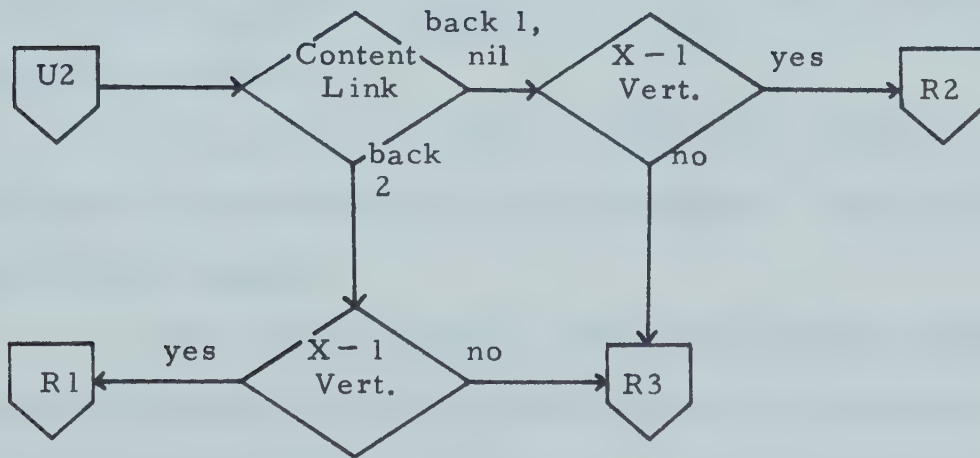
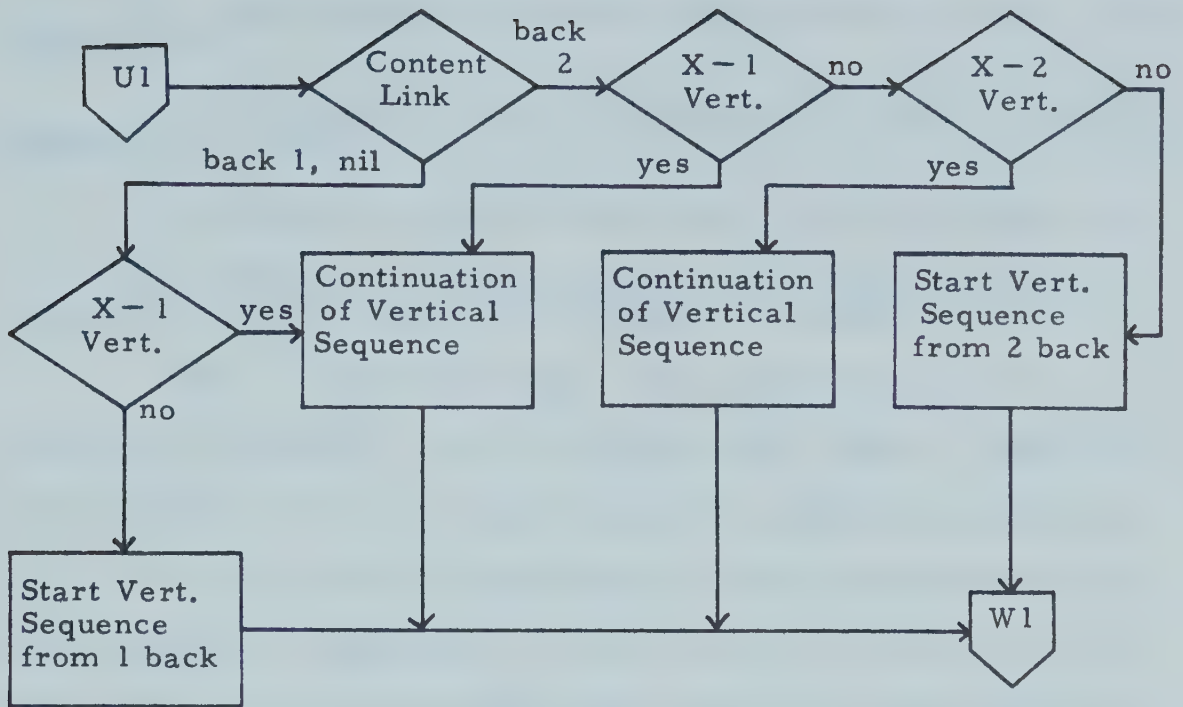


Figure 18: Continued



W. The rate of synthesis was expected to vary with temperature.

X. Examples of this relationship abound in natural systems, of course.

Y. However, such was not the case with Strain 32.

Z. Jeeves and Souse demonstrated that the rate of synthesis was invariant with respect to temperature.

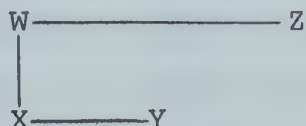
CASE 1: Were the above four sentences to be analyzed by PLATEXT, the following would result, ignoring any connections W and X may have with preceding sentences.

X: Sub x Nil → Sub Back 1. The subordinate relation indicator "EXAMPLES" is more heavily weighted in the dictionary of relation words than the coordinate indicator "OF COURSE". X and W have no content words in common, and we are ignoring pre-W sentences, so the content relation discovered is "Nil".

Y: Co x Nil → Co Back 1. The only relation indicator is the coordinate clue "HOWEVER". There are no content word matches.

Z: Nil x Nil → Super. There are no dictionary relation matches, and the sentences X and Y contain no content words in common with Z.

The resultant configuration for W, X, Y, and Z is



CASE 2: If the text offered the program consisted of the sentences W, X and Z, the following analysis would result.



X: Sub x Nil  $\rightarrow$  Sub Back 1. As in Case 1.

Z: Nil x Back 2  $\rightarrow$  Co Back 2. There are no dictionary relation words. Z has content words "rate", "synthesis", and "temperature" in common with W.

The resultant configuration for W, X, and Z is



CASE 3: If the text offered consisted of the sentences W, Y, and Z, the following analysis would have transpired.

Y: Co x Nil  $\rightarrow$  Co Back 1. As in Case 1.

Z: Nil x Back 2  $\rightarrow$  Co Back 2. As in Case 2.

The resultant configuration for W, Y, and Z is

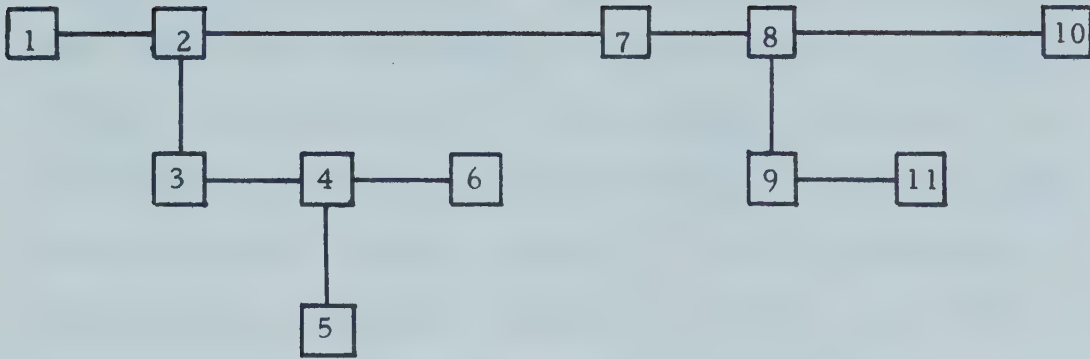


The pattern produced is a result of the interaction of the truth table components rather than a simple summation of them.

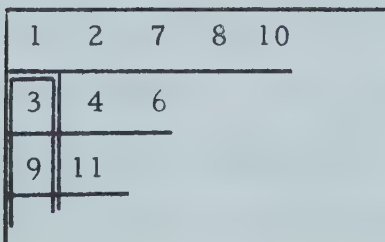
The purpose of the procedure outlined in flowcharts, Figure 18 (two parts), is to create the pattern in two matrices, and to provide an index to them. The scheme for representation is demonstrated in Figure 19. The simple 11-sentence text pattern diagram is shown with matrices and indexing array. HORZ and VERT are otherwise zero-filled. The double-lined areas of these contain sentences whose primary reference is in the other array. The references are compiled in INDX, where the nth element in the top row



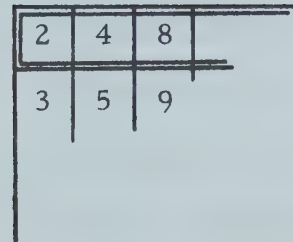




HORZ



VERT



INDX



1	1	2	1	2	1	1	1	2	1	1
1	1	2	2	2	2	1	1	2	1	3
1	2	1	2	2	3	3	4	3	5	2

Figure 19: Internal Representation



indicates which of HORZ and VERT (HORZ = 1, VERT = 2) contains the nth element. The remaining information in INDX gives the coordinates of each element in either HORZ or VERT, as appropriate. For example, Sentence 6 is described by the 6th column (pointer) in INDX as being located in HORZ, row 2, column 3. INDX contains no information not already available in HORZ and VERT, but in a form that the pattern-producing procedure can more readily use.

HORZ and VERT are designed to be easily transformable into the text pattern diagram; they serve as the interface to the plot production segment of PLATEXT. This segment is responsible for working through the elements of HORZ and VERT and queuing them in the appropriate sequence for the subroutine BEFIX. The pattern production procedure is outlined in flowchart, Figure 20, the objective of the algorithm being, for an n-sentence text, the proper sequence of n subroutine calls.

3.3.1.9 Example. This entire section may best be summarized by examining how PLATEXT analyzes a specific fragment of text. The sample text is an eight - sentence extract from a newspaper supplement article on the "chip" when such a device represented the latest generation of electronic marvels (40). Figure 21 contains the extract in its original order. Content words have been capitalized and dictionary relation words underlined. This tagging represents PLATEXT's first task: the isolation of terms



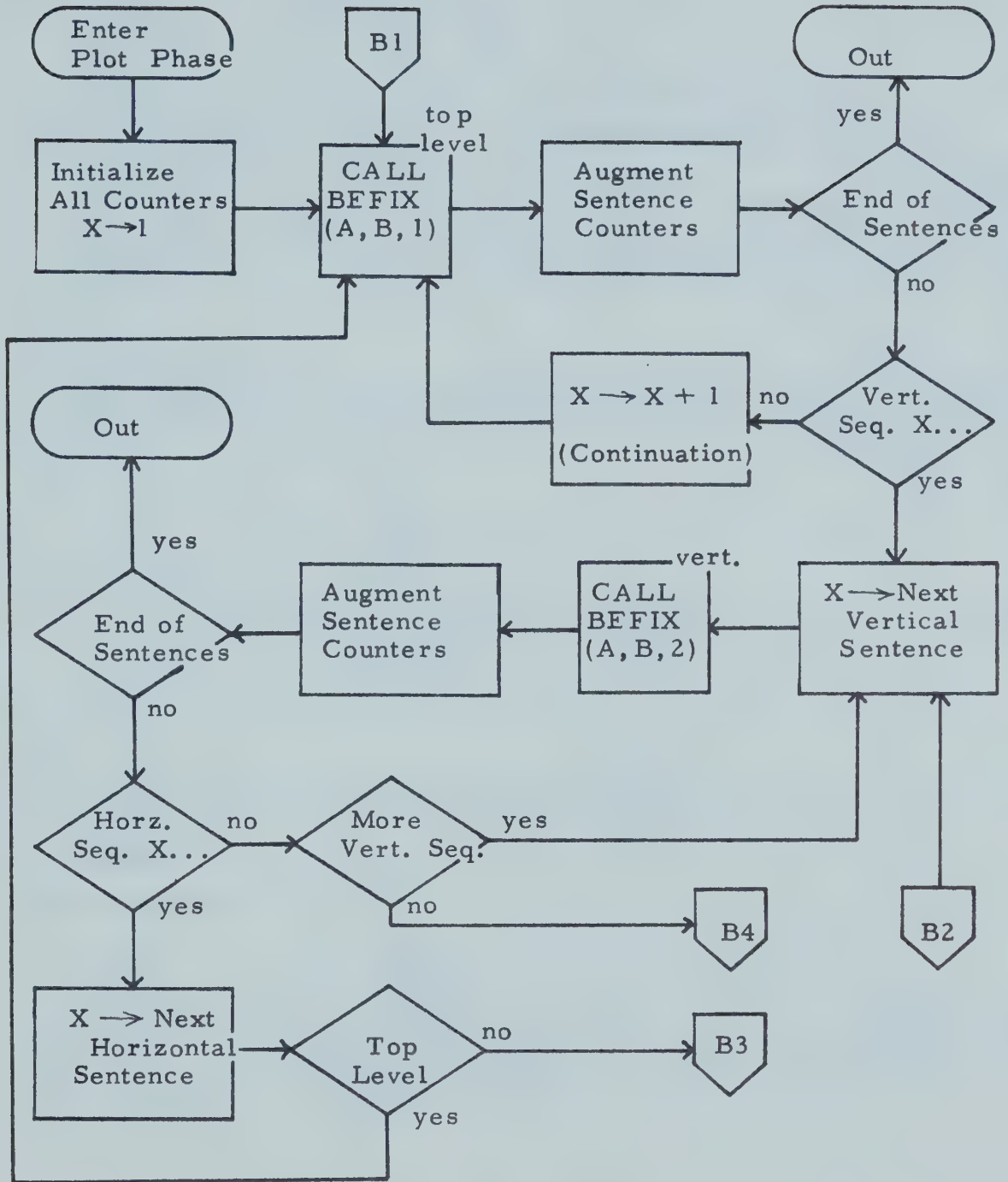


Figure 20: Plot Production Routine





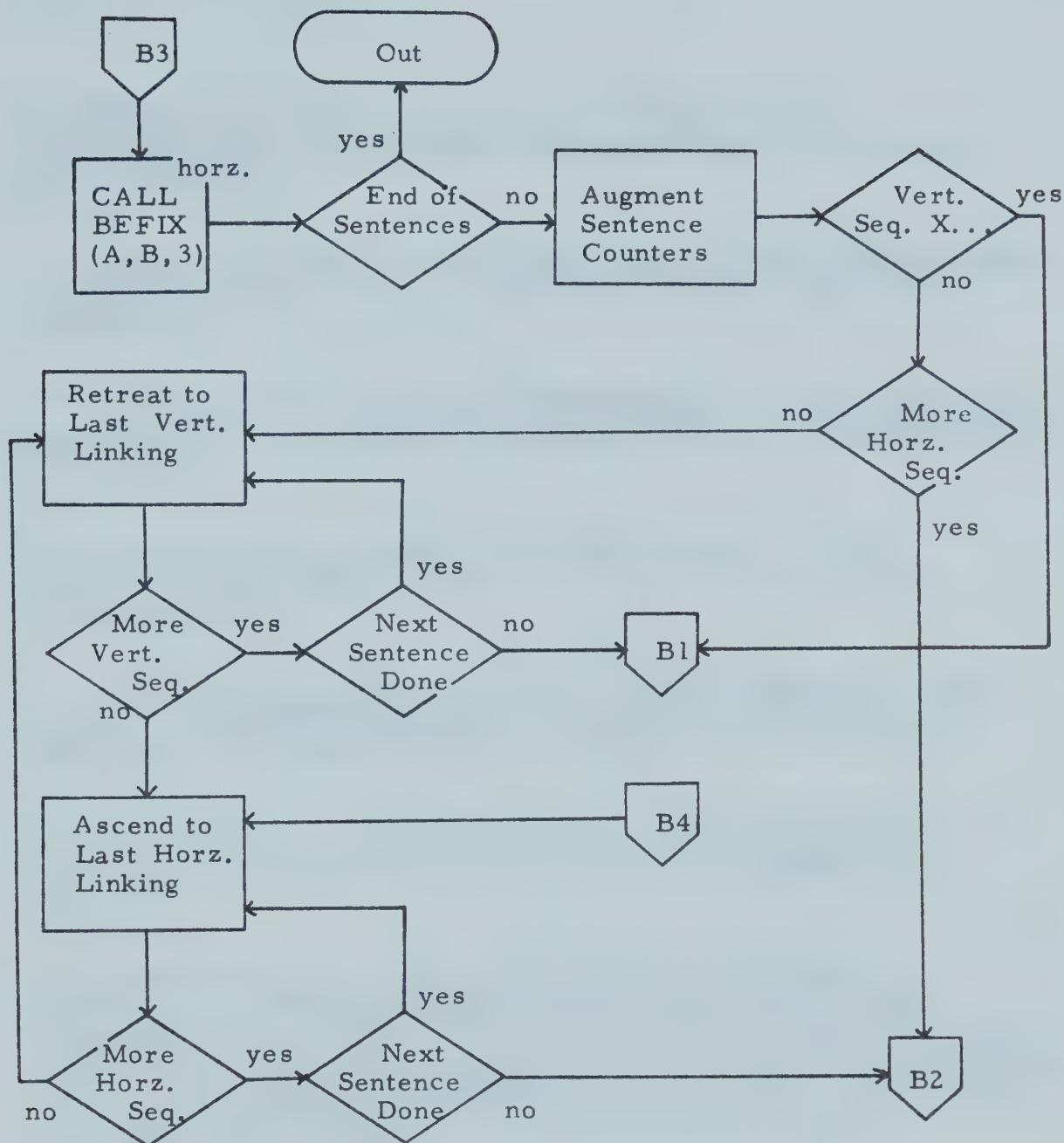


Figure 20: Continued



1. PRINTED CIRCUITS made the next CONTRIBUTION to MINIATURIZATION by ELIMINATING most of the WIRING between COMPONENTS.
2. in PLACE of WIRES small PATHWAYS of CONDUCTING MATERIAL are PRINTED DIRECTLY on a THIN board of NONCONDUCTING MATERIAL.
3. the TRANSISTORS and other COMPONENTS are then MOUNTED on these BOARDS, their CONNECTIONS SOLDERED to the CONDUCTING PATHWAYS.
4. the FINAL STEP in MINIATURIZATION -so far - is the INTEGRATED CIRCUIT, or CHIP, so CALLED because of its MINISCULE SIZE.
5. this is an ULTRA MINIATURE SOLID-STATE CIRCUIT in which not only the CONNECTIONS but the COMPONENTS themselves are INSEPARATELY ASSOCIATED.
6. in one process several HUNDRED TINY, IDENTICAL CIRCUITS are PRODUCED on a SINGLE SLICE of SILICON the SIZE of a QUARTER.
7. MICROSCOPICALLY small CHANNELS and OPENINGS are ETCHED in, and POSITIVE, then NEGATIVE MATERIALS are INTRODUCED in NEARLY INVISIBLE AMOUNTS, LAYER by LAYER and PATTERN by PATTERN, until COMPLETE MINUTE TRANSISTORS and other NECESSARY COMPONENTS are CREATED, each in PLACE and PROPERLY CONNECTED, within the SPACE of a few HUNDREDTHS of an INCH.
8. a HANDFUL of CHIPS can PROVIDE enough CIRCUITRY for a DOZEN COMPUTERS or THOUSANDS of RADIOS.

Figure 21: Sample; Original Order, with Dictionary and Content Tagging



meaningful for its purpose.

The distribution of the relation words over the 8 sentences of the text is stored in array IXRTC1 (Figure 22), and details of each of the 4 dictionary matches are given in IXRT1. The distribution of content word matches is given in IXRTC2, and details stored in IXRT2. PLATEXT then decides on a dominant category (one of twelve possible) for each sentence, and stores these choices in OVRL. Using OVRL, the program next creates HORZ, VERT, and INDX. The resultant pattern is shown in Figure 23 with the sentences in the order of plotting.

### 3.3.2 Subroutines: Subroutine BEFIX.

PLATEXT consists of a mainline program and eleven programmer-defined and library subroutines in the relationship diagrammed in Figure 24. The following sections describe each of the subroutines in a depth and manner corresponding to the previous discussion of the MAIN program of PLATEXT. The principal subroutine, and principal heading in this section, is BEFIX.

Subroutine BEFIX (A, B, C)

Local Arrays:	HOLD (2,50)	(REAL)
	KAUGV (50)	(INTEGER)
Arrays in COMMON:	HORZ (10,40)	(INTEGER)
	VERT (40,10)	(INTEGER)
Declared Variables:	A,B,C	(INTEGER)

Comments: BEFIX translates a specified element (sentence





HORZ

1 2 3 5 8

4 6 7

VERT

3

4

INDX

1 1 1 2 1 1 1 1

1 1 1 2 1 2 2 1

1 2 3 1 4 2 3 5

IXRTC1

0 0 0 1 2 0 1 0

IXRT1

1 1 2 2

1 1 1 1

IXRTC2

0 1 3 0 3 1 1 0

IXRT2

1 1 2 2 4 3 3 4 5

1 1 1 1 1 1 1 1 1

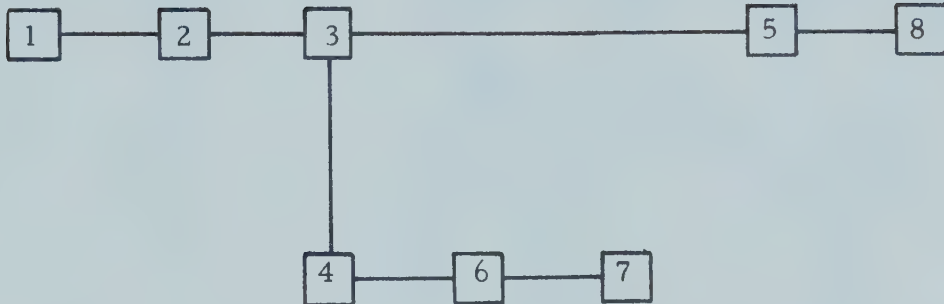
OVRL

4 4 2 1 2 4 2 4

3 1 1 3 2 2 2 3

Figure 22: Internal Representation





Printed circuits made the next contribution to miniaturization by eliminating most of the wiring between components. In place of wires small pathways of conducting material are printed directly on a thin board on nonconducting material. The transistors and other components are then mounted on these boards, their connections soldered to the conducting pathways. The final step in miniaturization - so far - is the integrated circuit, or chip, so called because of its miniscule size. In one process several hundred tiny, identical circuits are produced on a single slice of silicon the size of a quarter. Microscopically small channels and openings are etched in, and positive, then negative materials are introduced in nearly invisible amounts, layer by layer and pattern by pattern, until complete minute transistors and other necessary components are created, each in place and properly connected, within the space of a few hundredths of an inch. This is an ultra miniature solid-state circuit in which not only the connections but the components themselves are inseparately associated. A handful of chips can provide enough circuitry for a dozen computers or thousands of radios!

Figure 23: Pattern, and Sample in Revised Order



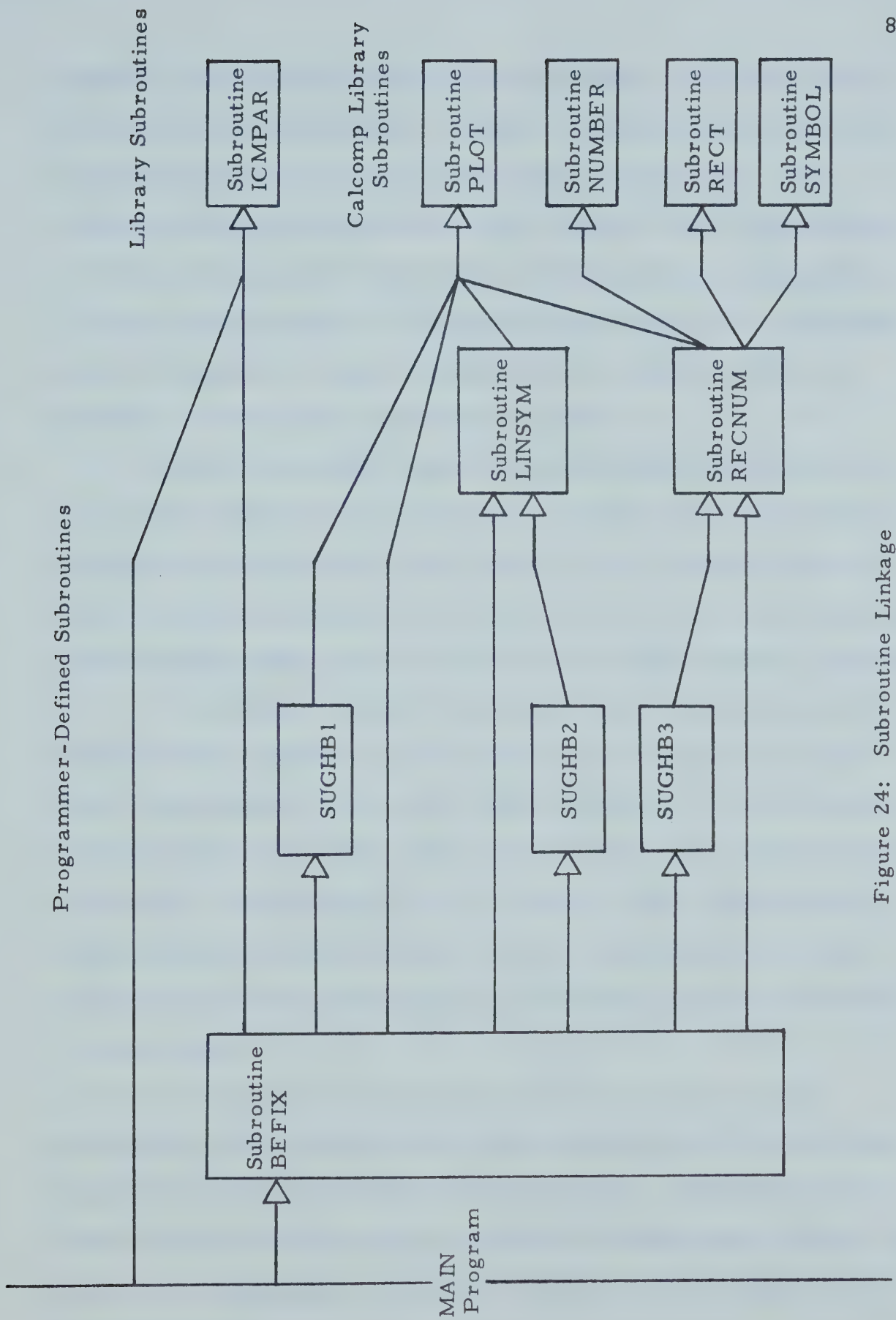


Figure 24: Subroutine Linkage





symbol) from the array HORZ (a,b) or VERT (a,b) into calls to the subroutines LINSYM, RECNUM, SUGHB1, SUGHB2, SUGHB3, and the CalComp subroutine PLOT (see Figure 24). The CalComp library subroutines PLOT, NUMBER, RECT, and SYMBOL, the last three called from the afore-mentioned programmer-defined subroutines as illustrated in Figure 24, ultimately place on magnetic tape individual pen commands for the creation of the text structure diagram.

BEFIX is called once for every sentence symbol to be plotted; these calls must be in succession. The responsibility of BEFIX is to keep track of the pattern drawn to date and the current pen position, so that the next sentence-identifying symbol can be placed in the appropriate position.

Structure: BEFIX is divided into 3 sections; one dealing with sentence symbols which are to appear on the first (principal) level, the second dealing with sentence symbols which are to be connected at the lower extremity of a vertical (subordinate) link, and the third dealing with those sentence symbols which are to be connected at the rightmost end of a horizontal (coordinate) link on any level but the first.

The first and second passed parameters in the subroutine call to BEFIX are the subscripts of the sentence symbol in either HORZ (a,b) or VERT (a,b). The sentence symbol to be plotted will be located in VERT only in the case when it is connected vertically, as one would expect from the appellation.



The last (third) passed parameter in the call to BEFIX is the indicator to which of HORZ and VERT and to which of the three sections of BEFIX control is to be passed. If  $C = 2$ , the subscripts A and B define element VERT ( $a = A$ ,  $b = B$ ) and the second section of BEFIX is to be executed. If  $C = 1$  or 3, A and B define element HORZ ( $a = A$ ,  $b = B$ ) and either the first or third section of BEFIX will be executed. In the case of sentence symbols on the first (principal) level ( $C = 1$ ), the specification of C is redundant since the first subscript (level indicator A) will be fixed at 1, i.e. HORZ ( $a = 1, b$ ).

Notes: Flowchart, Figure 27. 1. The variable LTRACK counts the number of horizontal links on all levels. Since the sentence symbol boxes and horizontal links are of uniform size whether they represent sentences and links that are superordinate or coordinate in nature, LTRCK can be used to measure the pattern's progress along the plot paper. To prevent wasting plot paper, when LTRCK equals 36 the pattern is restarted after a translation of -35.0 inches on the X-axis and -10.0 inches on the Y-axis. The translation is most easily accomplished, of course, if the 36 inch boundary is exceeded on the top level.

2. The array HOLDT contains the positional coordinates of the previous sentence symbol on the top level. This fixes one end of the required superordinate link.

3. The length of the graphic representation of the superordinate link is not fixed, it being a function of the



amount of textual material found by the analysis to be dependent on the sentence to which the current symbol is to be linked (see Figure 25). The variable HMAX is employed to keep track of the furthestmost extension of the pattern, and the horizontal positional co-ordinate HOLDT (1, NNOW) of the antecedent top-level sentence symbol is subtracted from the current value of HMAX to arrive at a "distance to be made up". In Figure 25a, HMAX and HOLDT(1,2) and HOLDT (1,3), HOLDT (1,4), and HOLDT (1,5) are set to the same value; no addition is necessary. In Figure 25b, HMAX will contain the value 2.0, and HOLDT (1,2) will be 1.0, indicating an additional 1.0 must be added to the link between sentence symbol 6 and its antecedent sentence symbol 2. In Figure 25c, HMAX = 3.0 so the link must be augmented by 2.0.

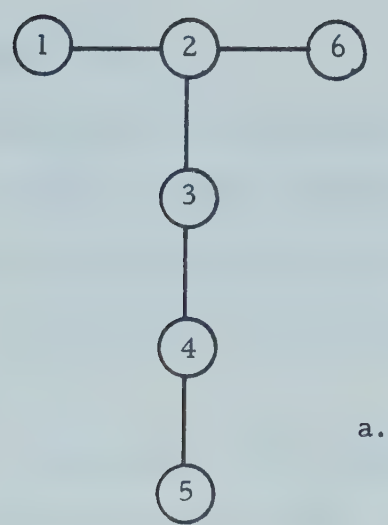
4. Before the link is drawn, i.e. before the subroutine LINSYM is called, a check is made on the value of LTRCK. If the value of LTRCK is less than 36, the link between the current sentence symbol and its antecedent is drawn. If this is the link that breaks the 36 unit barrier, the variables XRUPT and YRUPT are used in the translation back to the beginning of the second tier of the pattern. Once the second tier has been established, a second set of subroutines is employed in place of PLOT, RECNUM, and LINSYM.

5. The sentence symbol IDD is plotted at position PHORZ, PVERT, or if LTRCK is greater than 36.0, at

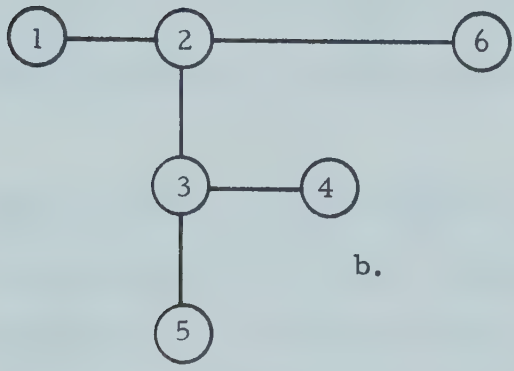




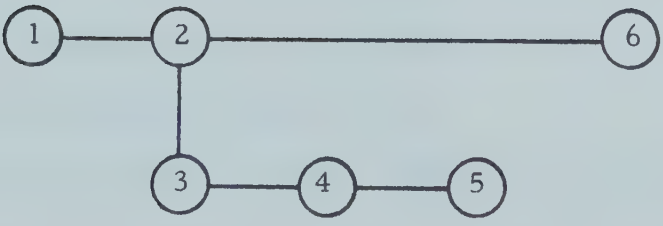




a.



b.



c.

Figure 25: Three Possible Linkages Between Sentence Symbol 2  
Sentence Symbol 6



PHORZ - 35.0, PVERT - 10.0.

6. The positional co-ordinates of sentence symbol IDD are stored in HOLDT (1,IDD) and HOLDT (2, IDD). HMAX is augmented, as it will be each time a superordinate or coordinate link is recorded, and control is returned to the mainline program.

Flowchart, Figure 28. The implementation of the algorithm in the case of the vertical (subordinate) link is simplified because, although many unconnected sentence symbols can reside in a given horizontal plane, only connected sentence symbols can reside in a given vertical plane (see Figure 26).

1. The initial step is to establish the positional co-ordinates of the antecedent sentence symbol. This is accomplished by referring to the array HOLDT. The index number (antecedent sentence symbol) is the element in VERT defined by A, B - 1. As was noted earlier, passing the subroutine BEFIX the parameters A, B, C with C = 2 (a vertical or subordinate link) implies that  $B \geq 2$ .

2. While the value of LTRCK cannot be augmented during the plotting of a vertical link the subordinate links are subject to the same "post-36" translation as horizontal links. As in flowchart, Figure 27, auxiliary subroutines are employed in place of PLOT, LINSYM, and RECNUM.

3. Several different schemes for the representation of relative dependency of sentences in text were tried during the development of PLATEXT. It was eventually decided



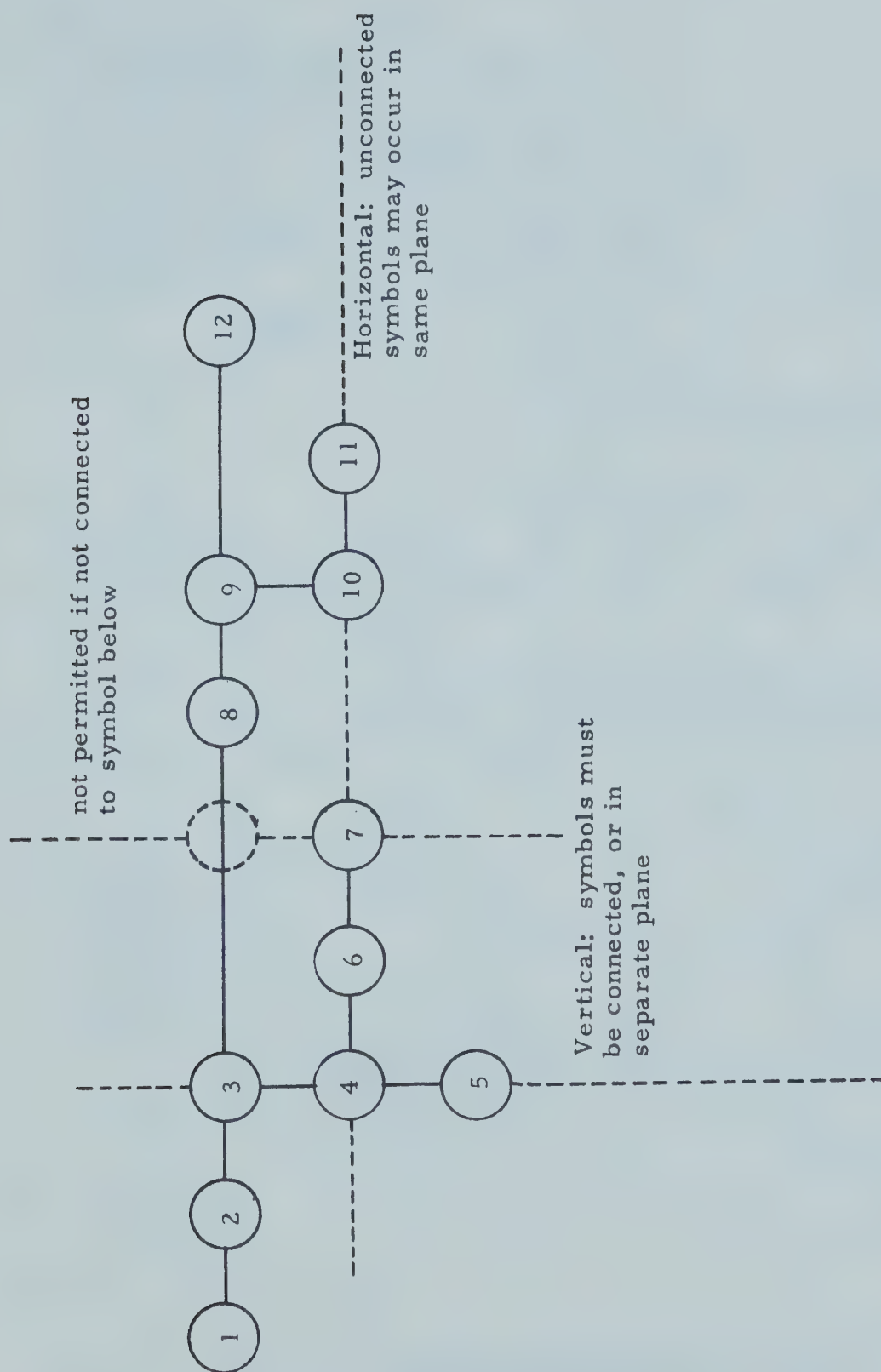


Figure 26: Sentence Symbol Assembly





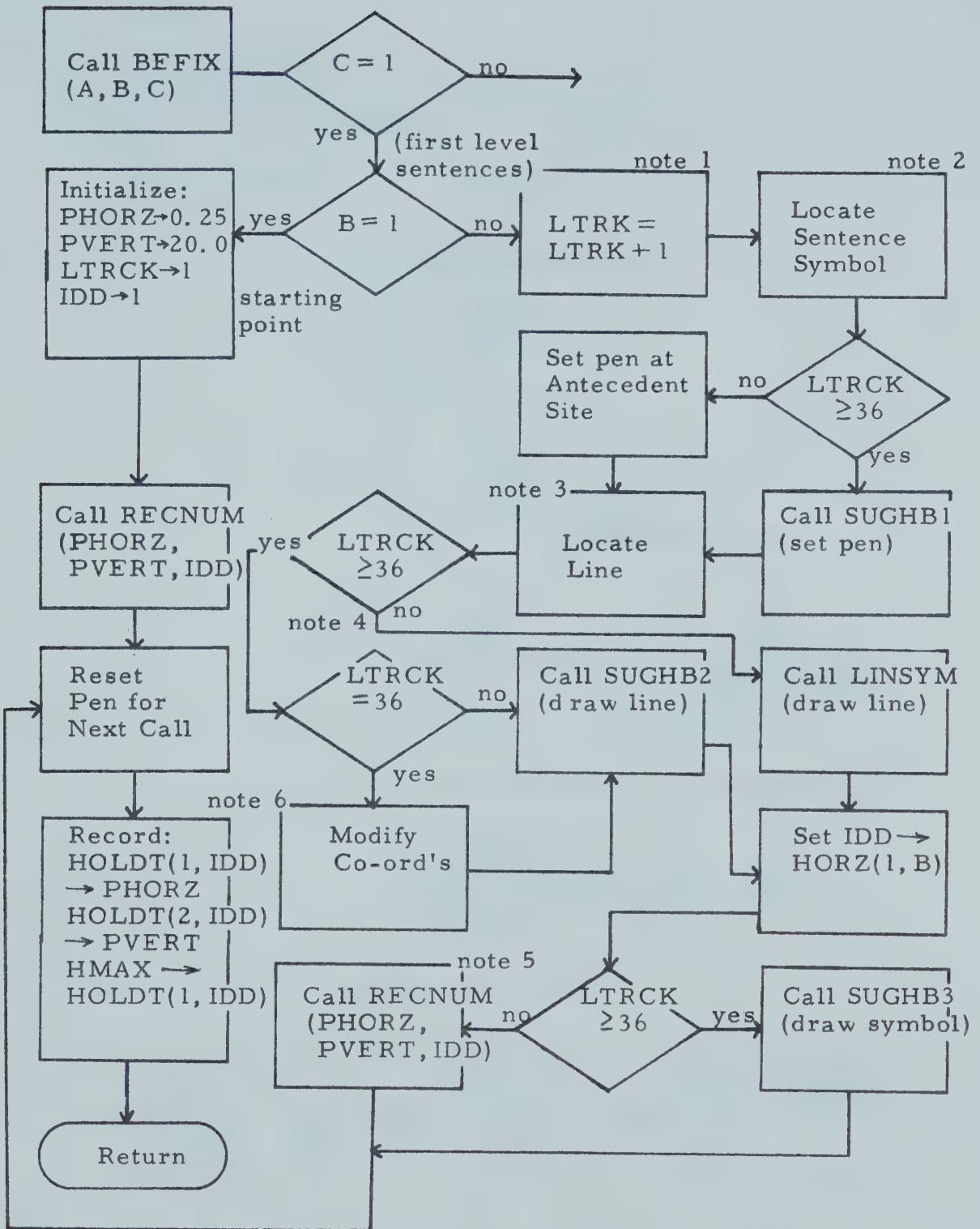


Figure 27: Subroutine BEFIX; Superordinate and Top-Level Coordinate Linkage



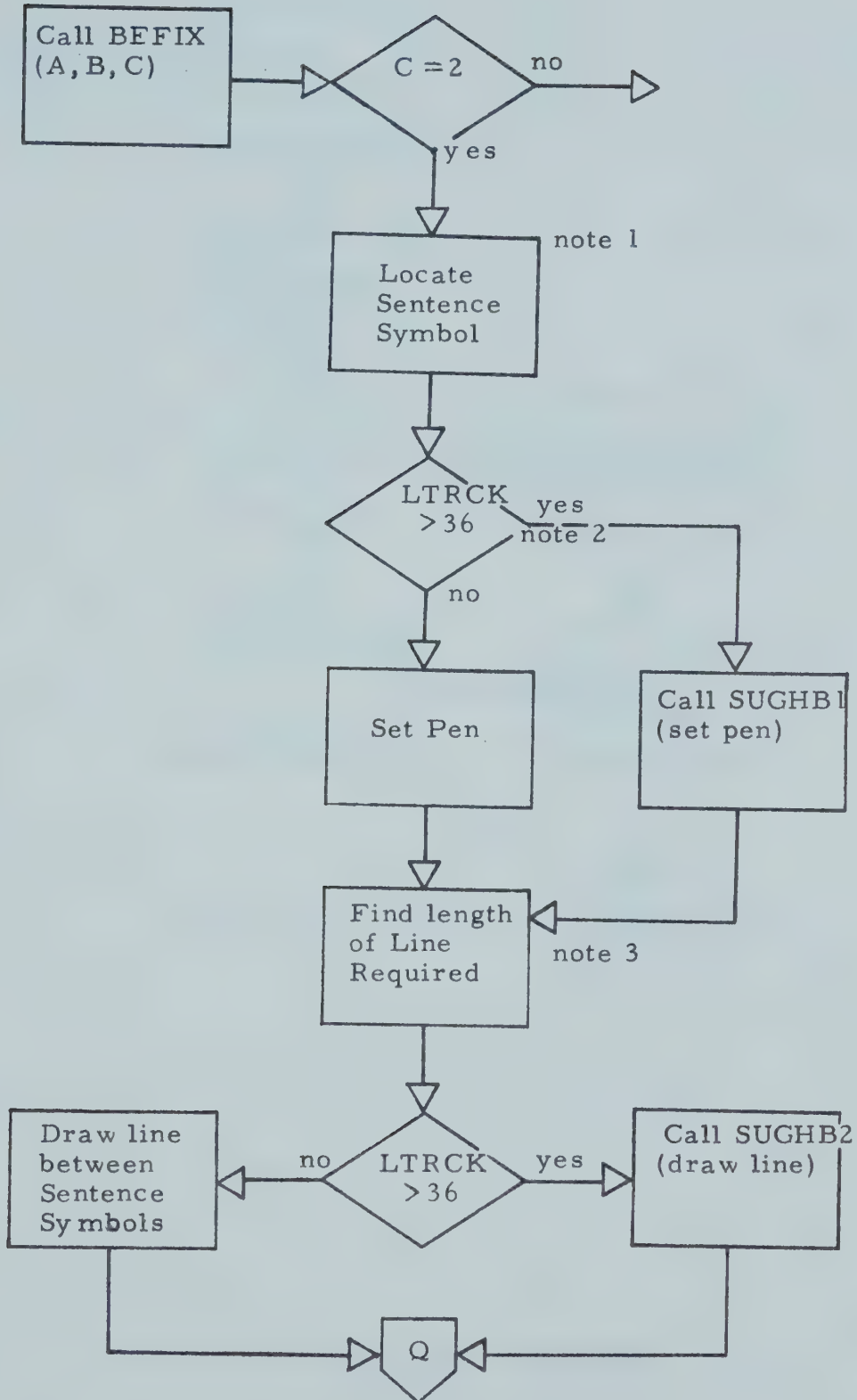


Figure 28: Subroutine BEFIX; Subordinate Linkage



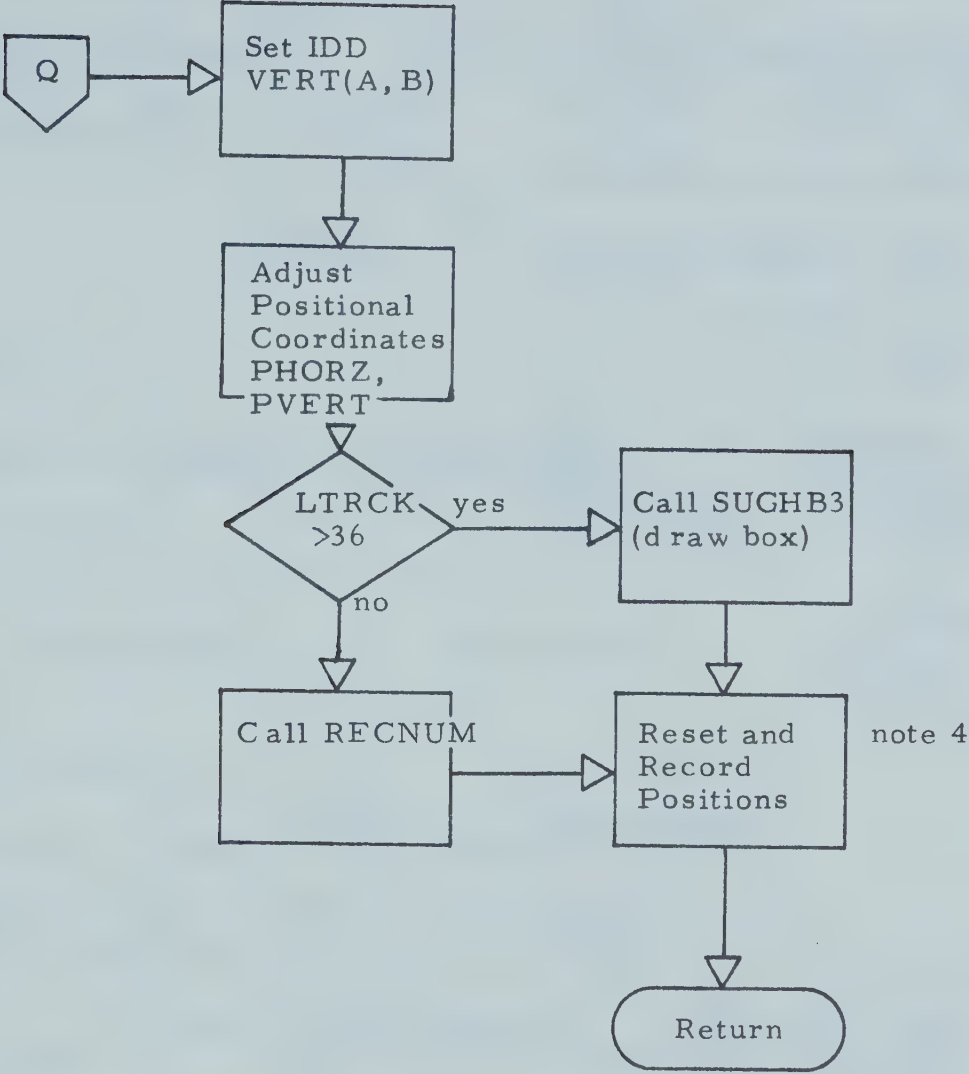


Figure 28: Continued





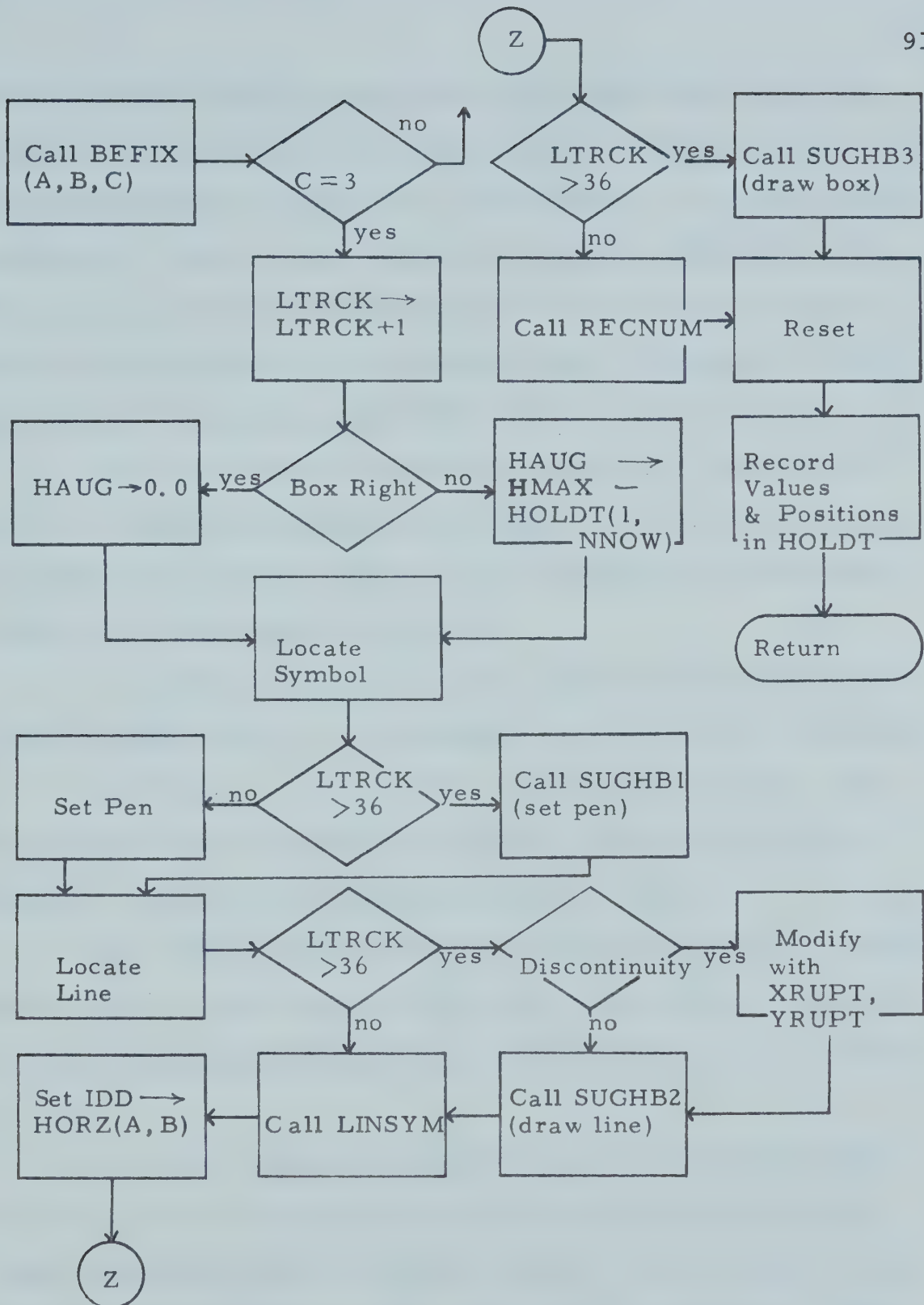


Figure 29: Subroutine BEFIX; Coordinate Linkage



that a fixed length of vertical link would best suit the model, but some of the variable-length link programming should be retained for the purpose of monitoring the pattern during analysis. Thus one has the array KAUGV(a) keeping track of the depth below the superordinate level of each sentence. HORZ (a,b) and VERT (b,a) contain this information as well, but not in an easily accessible form.

4. The link and sentence symbol are drawn with calls to LINSYM and RECNUM (if  $LTRCK \leq 36$ ), or with calls to SUGHB2 and SUGHB3 ( $LTRCK > 36$ ) analogously to the case when  $C = 1$  (flowchart, Figure 27).

Flowchart, Figure 29. The case of the subroutine call to BEFIX with  $C = 3$  indicates to the routine a horizontal (coordinate) link on other than the top level. The algorithm to be followed is basically that of the first segment of BEFIX ( $C = 1$ ; superordinate link).

1. The variable LTRCK is incremented by 1; the line drawn may in fact be much longer but the lateral increase in the pattern will always be 1 (for  $C = 1$  or 3) or 0 (for  $C = 2$ ).

2. The required length of the link is now established. The lateral positional co-ordinate of the antecedent sentence symbol (HOLDT(1, NNOW)) is subtracted from the maximum extension of the text structure pattern (HMAX) and the result (HAUG) is stored. Coordinate linking involves slightly more book-keeping than superordinate linking, as now the antecedent sentence symbol must be



located in a vertical plane. Again, one must check to see if the pattern has exceeded its horizontal boundary and is starting another tier.

3. The link is augmented by HAUG and drawn, subject to the results of tests for violation of the lateral boundary. There must be provision for the discontinuity in the pattern to occur on any level, but the procedure is identical to that discussed in the case of the top level.

4. The sentence symbol is drawn with calls to either RECNUM or SUGHB3, and the positional co-ordinates stored in HOLDT. HMAX is suitably modified.

#### 3.3.2.1 Subroutine RECNUM (A,B,C)

Passed Parameters:           A, B    (REAL)  
                                  C       (INTEGER)

Comments: The subroutine RECNUM draws a square 0.25 inches to the side at a site specified by the positional co-ordinates A, B. The integer number C is drawn inside the box. RECNUM calls the CalComp library subroutine RECT to accomplish the former, and either of the CalComp library subroutines NUMBER or SYMBOL to achieve the latter. NUMBER is called when  $C \geq 10$ ; SYMBOL is used when  $1 \leq C \leq 9$ , in which case the digits plotted are 01, 02, 03, .... These symbols (i.e. digit and leading zero) are stored in the character array NUMB.

The values passed to the subroutine RECNUM as A, B are the co-ordinates of the rightmost bottom corner of the square (X maximum, Y minimum). The subroutine RECT employs





as starting point the co-ordinates of the leftmost bottom corner (X minimum, Y minimum), necessitating a translation of 0.25 in the horizontal (X, or A) value. The same effect could have been achieved with a rotation of  $90^\circ$  (see Figure 30).

Various sizes of sentence symbols were tried, from 1 inch to the present 0.25 inch size, which seems to be the most compact compatible with readable sentence numbers; the digits are 0.10 inches tall.

### 3.3.2.2 Subroutine LINSYM (A,B,C,D)

Passed Parameters:            A, B, C, D            (REAL)

Comments: The subroutine LINSYM draws the lines connecting the sentence symbols. The parameters A, B are the positional coordinates of the start of the line, and C, D the coordinates of the end of the link. The actual drawing of the connector is accomplished by calls to the CalComp library subroutine PLOT, which could have been done from the subroutine BEFIX. However isolation of the link-drawing mechanism allows one to contemplate and easily insert instructions for the plotting of more complex and descriptive representations of links. In the past, PLATEXT put "little arrows" beside the links; a worthwhile modification might be an indication of the number of content links in the same location.

### 3.3.2.3 Subroutines SUGHB1, SUGHB2, and SUGHB3.

SUGHB1 (A, B)                      A, B                      (REAL)

SUGHB2 (A, B, C, D)      A, B, C, D      (REAL)



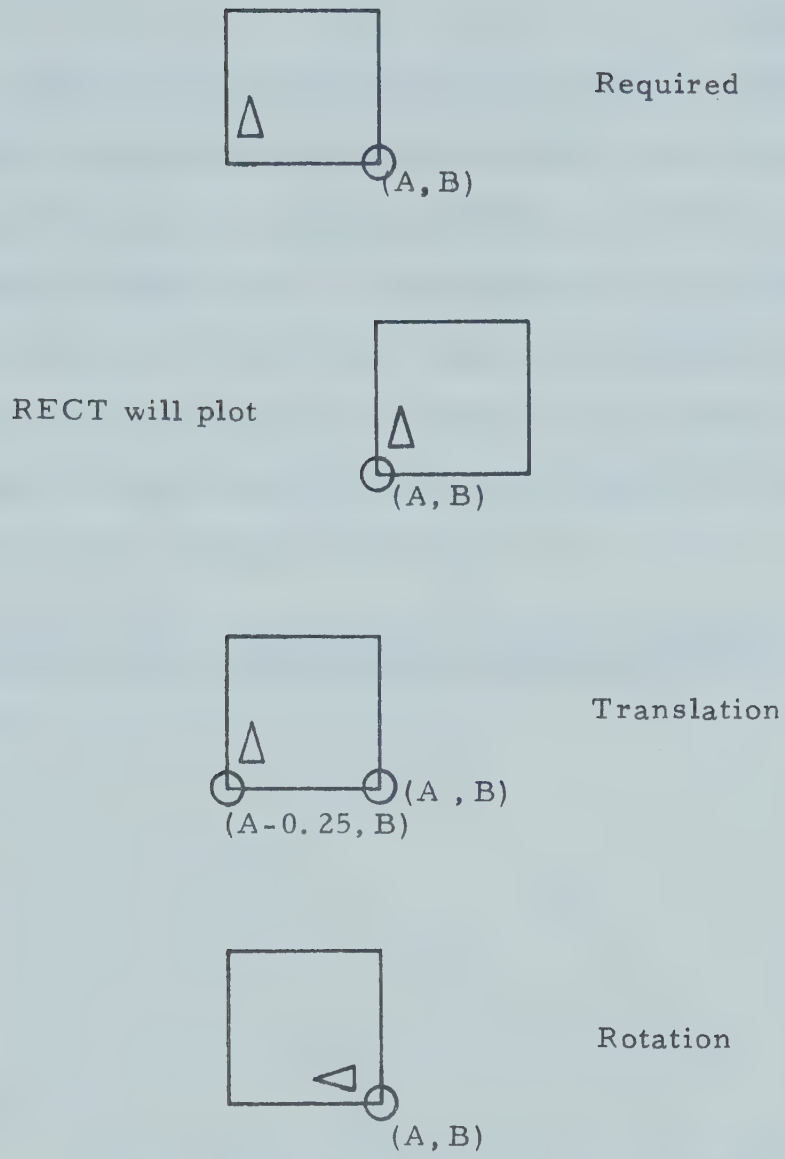


Figure 30: Rectangle; Orientation



SUGHB3 (A, B, C)	A, B	(REAL)
	C	(INTEGER)

Comments: These three subroutines are called instead of PLOT, LINSYM, and RECNUM respectively when the lateral extension of the text structure pattern exceeds 36 inches. In each case a translation of co-ordinates is performed ( $X_1 = X_0 - 35.0$ ,  $Y_1 = Y_0 - 10.0$ ) and PLOT, LINSYM, or RECNUM called. By moving this translation to subroutines, the machinery in BEFIX which tracks the pattern's progress is kept relatively straight-forward.

### 3.3.3 CalComp Library Subroutines Used in PLATEXT

PLOTS (A, B)

PLOT (C, D, E)

NUMBER (F, G, H, I, J, K)

SYMBOL (L, M, N, O, P, Q)

RECT (R, S, T, U, V, W)

C, D, F, G, H, I, J, L, M, N, P, R, S, T, U, V (REAL)

A, B, E, K, Q, W (INTEGER)

Comments: 1. PLOTS (A,B) This subroutine specifies a buffer area of size B with starting address A. This storage space is used in, among other things, I/O operations for plotter commands and initiates the plot program.

2. PLOT (A, B, C) This subroutine causes the plotter pen to be moved from its current location to that specified by co-ordinates A, B. If  $C = 3$ , the pen is in a raised position during transit (no line); if  $C = 2$ , the pen





is lowered and a line drawn from the pen position at the time of the call to PLOT to position, A, B.

A call to PLOT with  $A = 0.0$ ,  $B = 0.0$ , and  $C = 999$  causes plotting to be terminated. A block address of 999 is written on the tape, and the plottape file is closed.

3. NUMBER(A, B, C, D, E, F) This subroutine causes the plotter commands necessary to draw the floating point number D to be placed on the plottape. The number will be drawn at location A, B and each digit will have a height of C inches. E is the number of degrees the number is to be rotated, if desired. F specifies a number of modes of operation; for the purpose of PLATEXT, F is kept at -1, indicating to the subroutine NUMBER that only the integer portion of D should be plotted.

4. SYMBOL(A,B,C,D,E,F) This subroutine can be used to plot any of 128 symbols, including letters, digits, and most useful special characters. The items to be drawn can be placed in an array, e.g. NUMB in subroutine RECNUM, and referenced, or the index number of the desired symbol can be specified if known beforehand. The former procedure is employed by PLATEXT, with the characters 01XX, 02XX, 03XX, . . . stored in the array NUMB. The characters are drawn at position A, B with height of C inches and rotation of E degrees. F tells the subroutine how many characters of the array to plot, starting at location D. (In the PLATEXT application, F is set at 2.)

5. RECT (A, B, C, D, E, F) This subroutine causes



pen commands sufficient to draw a rectangle of height C and width D to be placed on the plottape. The rectangle will be drawn at the position specified by co-ordinates A, B with a rotation of E degrees. F is either 2 or 3 indicating whether the pen should be in a lowered or raised position as it moves from its present position to the start of the rectangle. PLATEXT employs 3.

A possible application for the rotation parameter, which measures the angle made by the rectangle base with the horizontal (Figure 32) is discussed in the section dealing with subroutine RECNUM.

### 3.3.4 Library Subroutine ICMPAR\*

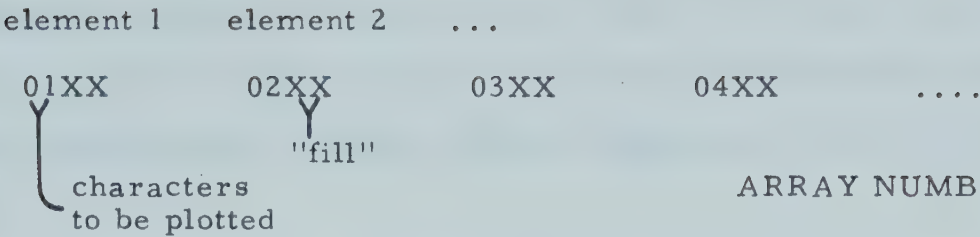
This Computing Centre Public Library subroutine was written by Mr. F. Jacobson in Assembler 360, and is employed by PLATEXT to make micro-second comparisons between strings of FORTRAN characters. The character strings to be compared must be of equal length, and not over 256 characters long. The general form of the subroutine call is CALL ICMPAR (FIRSTSTRING, SECONDSTRING, LENGTH, INDICATOR) where FIRSTSTRING and SECONDSTRING are the character strings to be compared, LENGTH is their mutual length in characters, and INDICATOR is the parameter which will be set to +1, 0, or -1 depending on the relationship between FIRSTSTRING and SECONDSTRING.

---

\*

Now CS286A under MTS.





CALL NUMBER (A, B, C, NUMB(3), E, 2) causes "03" to be plotted at A, B with height C and rotation E.

Figure 31: Example; CALL NUMBER ( A, B, C, NUMB(3), E, 2)

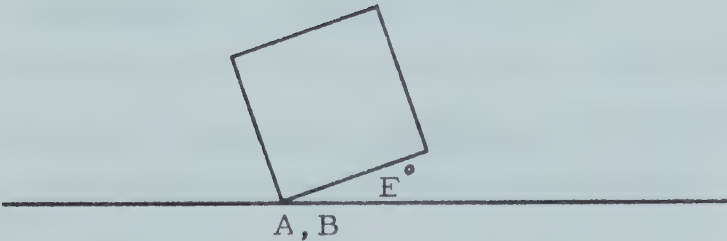


Figure 32: Rotation Parameter





The requirements of ICMPAR are responsible for the word length tabulating incorporated in various parts of PLATEXT. Each word in the text must undergo at least one comparison against some or all of the dictionary, against some or all of the erasure list, and perhaps some content word comparisons; ICMPAR effects substantial savings in time.

Substantial savings in core could have been effected had an assembler routine to "pack and unpack" FORTRAN character strings been available. The text handling and storage performed by PLATEXT is done on characters whose representation is X ~~øøø~~ (one character and three blanks per word), admittedly not very efficient. While the manipulation of individual characters will always be conducted in such a mode, as long as FORTRAN and not PL/1 is used in PLATEXT, much of the storage and manipulation of units larger than a character could take place in "packed" form (4 characters per word). All that is lacking is a suitable assembler routine for converting back and forth.

### 3.4 General Procedure for Augmented Text

In addition to having headings, numbering, captions, etc. (see section on General Procedure for Unedited Text for details of input requirements) deleted from the sample text, augmented text would have a part-of-speech (PS) code appended to each word (Figure 33). The projected PS codes were:



THE7 INTEGRATING3 CHARACTERISTICS1 OF6  
THE7 NEURON1 ARE2 NOT9 COMPLETELY4 DESCRIBED2 BY6  
THE7 ASSERTION1 THAT8 INCOMING3 SIGNALS1 PRODUCE2  
DEPOLARIZING3 EFFECTS1 WHICH8 ARE2 SUMMED2 IN6  
THE7 BODY1 OF6 THE7 CELL1.

Figure 33: Sample of Augmented Text



1. Nouns
2. Verbs and verb forms
3. Adjectives
4. Adverbs
5. Pronouns
6. Prepositions
7. Articles
8. Conjunctions
9. Others

The advantages foreseen for the employment of augmented text were threefold: firstly, the dictionary of relation words could be made more specific, perhaps with entries keying on combinations of PS codes and words. For example, "HOWEVER, THE BOTTOM FILE . . ." and "HOWEVER TALL THE SUBJECT . . ." illustrate two uses of the word HOWEVER with, presumably, two differing connective roles. A PS code could form the basis for distinguishing between them. Secondly, content matching could be made much more exact. Securing this advantage was the primary motivation in investigating the augmented text approach. Typically, the procedure would seek an antecedent noun for each pronoun encountered in the text. This would establish a demonstratable and reliable content link between the sentences involved. Thirdly, more meaningful and comprehensive text statistics could be compiled.

The disadvantages were obvious. A tremendous amount of effort is needed to accurately pre-edit the text to





provide the required information for the PS codes. A computer program might be developed to provide the codes, but such a project would eclipse the present one. The analysis of the text into parts-of-speech is the sort of task that people still perform much more economically than machines, but even so, manual pre-editing is prohibitively expensive. Nor will manual or automated procedures likely be able to economically provide the required analysis in the near future.

The only difference in the actual procedures in the analysis of unedited and augmented text was in the method of the establishment of content links between sentences. As previously mentioned, in the manual analysis a directed search was instituted, as well as the normal content word matching as described in the sections dealing with the program PLATEXT. The directed search attempted to find nouns to match the pronouns encountered in the current sentence. In a computer implementation to accomplish this, it would be necessary to open the boundaries of the matching procedure and let the directed search mechanism scan the entire text rather than a two-sentence radius as in the content word matching scheme of PLATEXT.

As stated, the preliminary and intermediate results with augmented text were not at all encouraging. Firstly, very few dictionary entries profited by the extra specificity obtainable with augmented text. The alternate scheme of weighting the dictionary entries for individual reliability



performed much of the adjustability originally envisaged for the PS codes - dictionary entry combinations. Secondly, the directed search, as tested manually, proved to be infrequently effective and very often inaccurate. The mere matching of nouns and pronouns is insufficient. Account must be taken of number, gender, and plausibility. No automated system in existence can do this with adequate precision. Thirdly, the lack of really comprehensive text statistics was regretted, but was deemed tolerable for the purposes of the project.

Finally, the most effective argument against the further development of the augmented text version of the analysis was the promise shown by the much less complex and less costly unedited text version of the analysis.



## CHAPTER IV

### SAMPLES OF TEXT

This section will discuss in detail the results of the analysis of one text sample, and will make general comments on the analyses of other samples included or referred to in Appendices C and D.

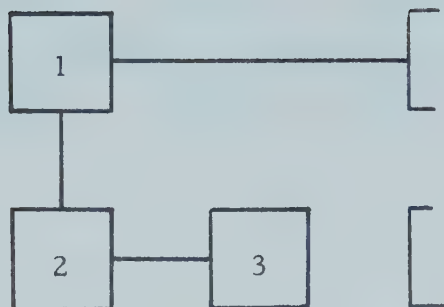
#### 4.1 Extract from The Genetic Code

The sample text for detailed consideration consists of thirty-one sentences from The Genetic Code by Isaac Asimov (41). The text structure diagram for this sample is given in Figure 40 in this section, and again as Figure C-10 in Appendix C. The text of the sample arranged in the pattern established by PLATEXT is displayed in Figures 34 to 39. The following sentence-by-sentence commentary will be concerned for the most part with the substructures in the diagram: those sixteen sentences which find themselves beneath the top level.

- |      |   |
|------|---|
| 1    | Sentence 1 is automatically placed on the top level.  |
| 2, 3 | Sentence 2 is subordinated because of the indicator PERHAPS, and Sentence 3 is linked to it by the word BREAKTHROUGH. |
| 4    | Sentence 4 is restored to the top level because of the absence of links. However,                                     |



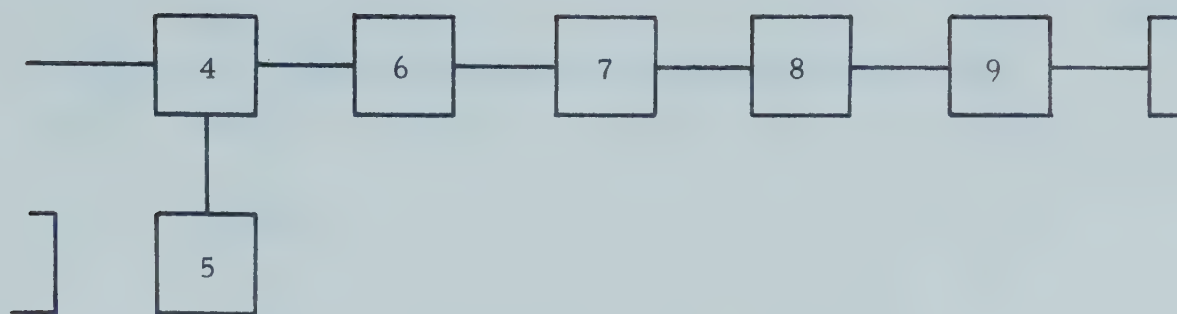




- 1 There still remains the question of the actual key of the code: which triplet stands for which amino acid?
- 2 The first breakthrough in this direction came in 1961 in what was perhaps the most important advance since the Watson-Crick model was proposed eight years earlier.
- 3 The breakthrough was the result of an experiment by Marshall W. Nirenberg and J. Heinrich Matthaei at the National Institutes of Health

Figure 34: Sentences 1-3

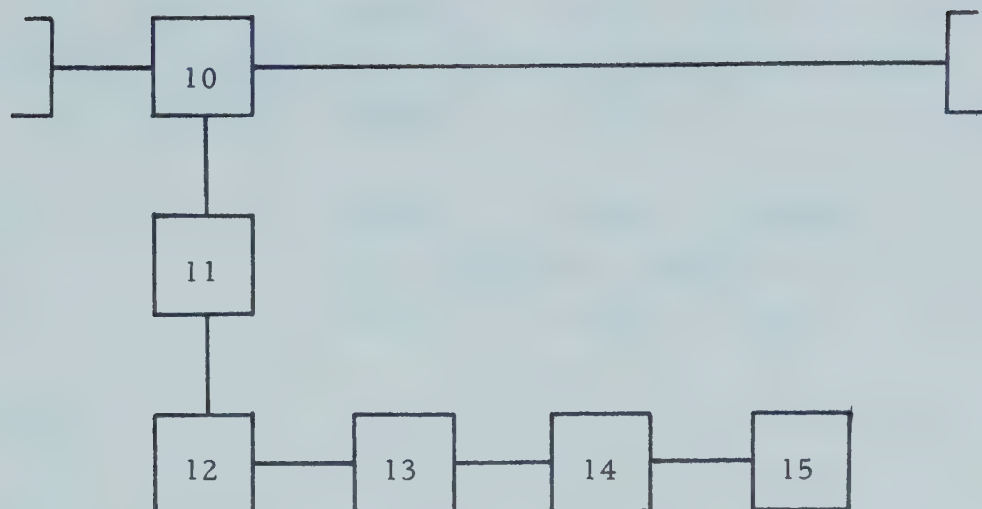




- 4 They realized that in order to learn the key, it was necessary to start with the simplest possible situation -- a nucleic acid made up of a chain of one single variety of nucleotide.
- 5 Ochoa had already shown how such a chain could be built up with the help of the proper enzyme, so that polyuridylic acid, for instance, could be easily manufactured and used.
- 6 Nirenberg and Matthaei therefore added polyuridylic acid to a system that contained the various amino acids, enzymes, ribosomes, and all the other components necessary to synthesize proteins.
- 7 Out of that mixture tumbled a protein that was as simple as the RNA they had in the beginning.
- 8 Just as the nucleic acid was all uridylic acid, so the protein was all phenylalanine.
- 9 This was important.

Figure 35: Sentences 4-9



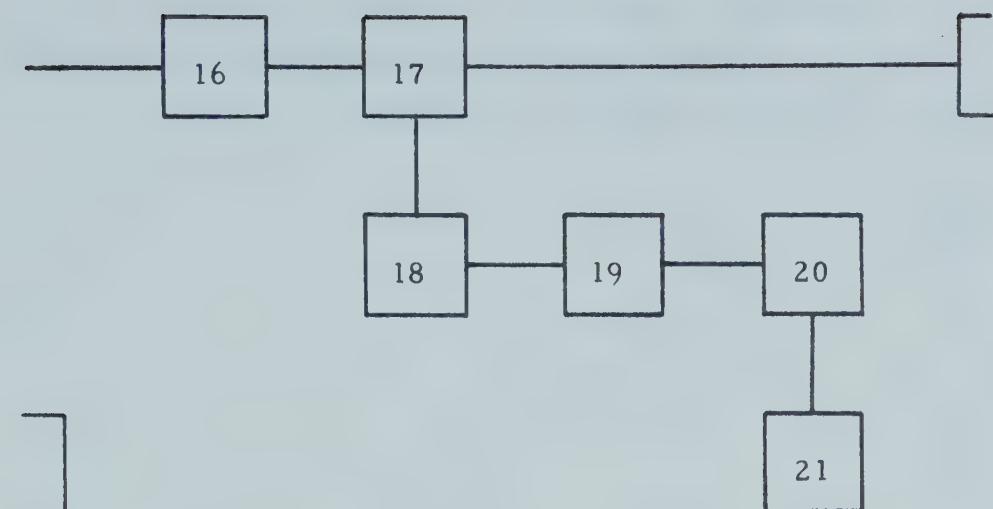


- 10 Polyuridylic acid could be represented as UUUUUUUUUUUU....
- 11 The only possible triplet that can exist in such a chain is, of course, UUU.
- 12 The only amino acid used in building the polypeptide chain was phenylalanine, although all the different amino acids were present and available in the system.
- 13 The conclusion that can be drawn from this is that the triplet UUU is equivalent to the amino acid phenylalanine.
- 14 The first step had been taken toward the decoding of the genetic code: "UUU means phenylalanine " was the first item in a "triplet dictionary ".
- 15 The next step was siezed upon at once; a number of research groups swung into action, following the lead that had been given them.

Figure 36: Sentences 10-15



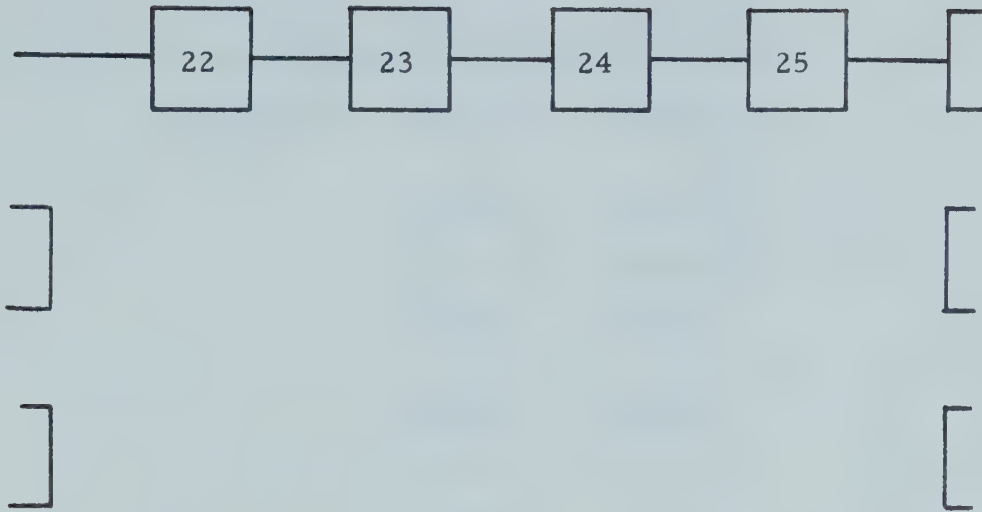




- 16 Suppose a polynucleotide is built up enzymatically out of a solution of uridylic acid to which a little adenylic acid has been added.
- 17 The chain will consist mostly of U, with an occasional A appearing at random.
- 18 The chain may then be, for instance, UUUUUUUUUUAUUUUUUUUUAUUUUUUUAUUUU....
- 19 Such a chain would be made up of the following triplets: UUU, UUU, UUU, AUU, UUU, UUU, UAU, UUU, UUA, UUU....
- 20 The triplets are still for the most part UUU, but occasionally an AUU, UAU, or UUA will creep in.
- 21 (These are the only three triplets that can be built from two U's and an A).

Figure 37: Sentences 16-21

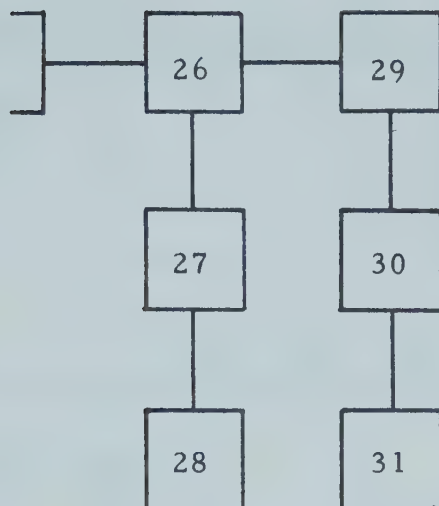




- 22 Sure enough, the protein formed by such an "impure " polyuridylic acid turned out to be mainly phenylalanine, but with occasional "intrusions " of other amino acids.
- 23 Three such "intruders " have been detected: leucine, isoleucine, and tyrosine.
- 24 It seems clear that one of the three triplets AUU, UAU, or UUA stands for leucine, one for isoleucine, and one for tyrosine.
- 25 Which is which, however, has not, at the moment of writing, been decided.

Figure 38: Sentences 22-25





- 26 The best we can do is write UUA in parentheses (UUA), and permit that to signify the three different triplets that can be built from two U's and an A, without even trying to specify the order.
- 27 In that case our dictionary could read: "(UUA) means leucine, isoleucine, or tyrosine".
- 28 If instead of adenylic acid, a little cytidylic acid or a little guanylic acid is added to the original solution of uridylic acid, polynucleotides are built up containing triplets that are (UUC) and (UUG).
- 29 Again, the parentheses mean that we are not specifying the exact order of the three nucleotides.
- 30 In both these latter cases, leucine can still be detected in the still chiefly phenylalanine protein that is produced.
- 31 This can only mean that (UUA), (UUG), and (UUC) can all be translated as leucine -- an example of what we have called the "degeneracy" of the code.

Figure 39: Sentences 26-31





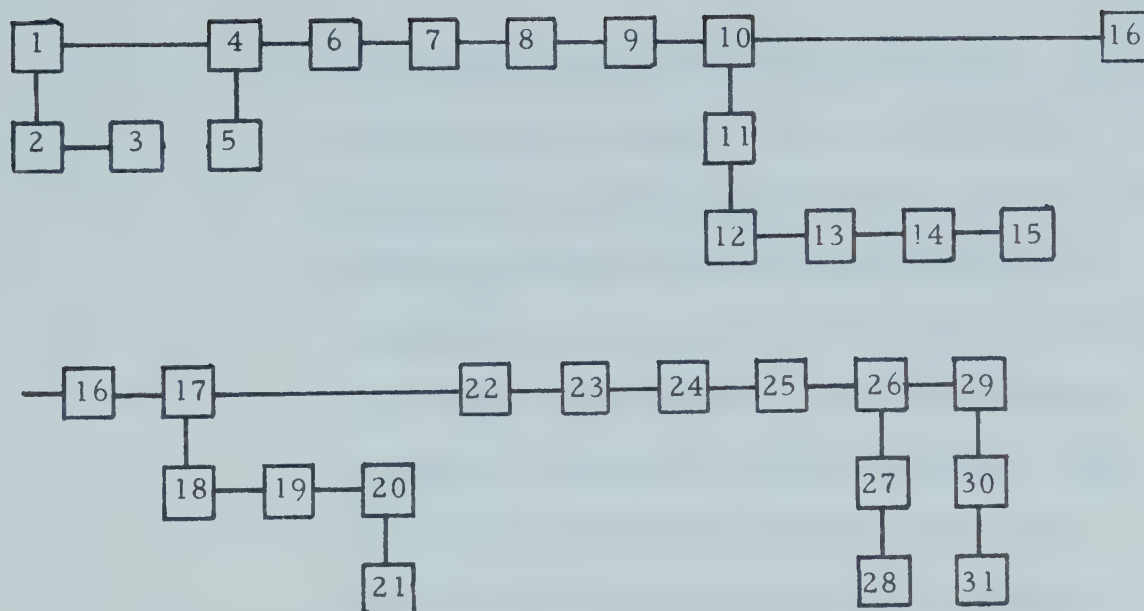


Figure 40: Text Diagram; Sample from The Genetic Code

There still remains the question of the actual key of the code: which triplet stands for which amino acid? They realized that in order to learn the key, it was necessary to start with the simplest possible situation — a nucleic acid made up of a chain of one single variety of nucleotide. Nirenberg and Matthaei therefore added polyuridylic acid to a system that contained the various amino acids, enzymes, ribosomes, and all other components necessary to synthesize proteins. Out of that mixture tumbled a protein that was as simple as the RNA they had in the beginning. Just as the nucleic acid was all uridylic acid, so the protein was all phenylalanine. This was important. Polyuridylic acid could be represented as UUUUUUUUUUUU.... Suppose a polynucleotide is built up enzymatically out of a solution of uridylic acid to which a little adenylic acid has been added. The chain will consist mostly of U with an occasional A appearing at random. Sure enough, the protein formed by such an "impure" polyuridylic acid turned out to be mostly phenylalanine, but with occasional "intrusions" of other amino acids. Three such "intruders" have been detected: leucine, isoleucine, and tyrosine. It seems clear that one of the three triplets AUU, UAU, or UUA stands for leucine, one for isoleucine, and one for tyrosine. Which is which, however, has not, at the moment of writing, been decided. The best we can do is write UUA in parentheses (UUA), and permit that to signify the three different triplets that can be built up from two U's and an A, without ever trying to specify the order. Again, the parentheses mean that we are not specifying the exact order of the three nucleotides.

Figure 41: Text Summary; Sentences from Top Level of Figure 40



- if PLATEXT examined three sentences previous for content links instead of two, Sentence 4 would be placed here due to its strong content linking with Sentence 1.
- 5 Sentence 5 is subordinated because of the occurrence of INSTANCE. As the diagram suggests, Sentence 5 may be removed from the text without disrupting the main theme, or, in this case, even disturbing it.
- 6 Sentence 6 is restored to the top level because of the occurrence of THEREFORE.
- 7, 8, 9, 10 This string is held to the top level by either content linking or a lack of links of any kind (Sentence 9).
- 11 This is subordinated through ONLY and OF COURSE.
- 12 This is subordinated through ONLY and ALTHOUGH.
- 13, 14, 15 Sentence 13 is coordinate-linked through THIS, and content-linked to Sentence 11 through TRIPLET and UUU. Sentence 13 is also content-linked to Sentence 12 through AMINO, AMINO, and ACID, which is a stronger link. Sentence 14 has no relation indicators in it, but strong content links to the sentences previous; a pity, because one suspects this sentence should be on



the upper level. Sentence 15 has no indicators as well, and is content-linked to Sentence 14. Sentences 14 and 15 are obviously a unit; where Sentence 14 is placed should determine Sentence 15's position.

16 Sentence 16 admits no links, so is placed back on the top level. In the original text, the paragraph break was between Sentences 14 and 15, with Sentence 16 being the second sentence in the paragraph initiated by Sentence 15. The arrangement suggested by PLATEXT that the paragraph start with Sentence 16 seems more logical.

17 Sentence 17 has no links, and so remains on the top level.

18,19,20,21 Sentence 18 is subordinated because of the occurrence of the indicator word INSTANCE. Sentences 19 and 20 are content-linked to Sentence 18 through the words CHAIN, TRIPLETS, UUU, UUA, etc. Sentence 21 is a still more dependent sentence, and, like Sentence 6, can be removed from the text without disturbing the main theme, or in this case, the subtheme 18, 19, 20.

22, 23,24,25,26 Sentences 22 through 26 are linked through





- content and occasional coordinate links.
- 27, 28      The phrase IN THAT CASE subordinates Sentence 27, and Sentence 28 is subordinated to 27 because INSTEAD (subordinate indicator) has a greater weight than IF (coordinate indicator).
- 29          Sentence 29 contains no links, so it is restored to the top level.
- 30, 31      This is a repeat of the situation of Sentence 27 and 28. The subordinate indicators involved are LATTER, ONLY and EXAMPLE.

A rough test of the effectiveness of the text structure-seeking procedures in PLATEXT is an examination of the sentences in the substructures, and a read-through of the sentences on the top level. If the sentences in the substructures seem to be concerned mainly with examples (particularly isolated cases such as Sentence 6) and reasonably consistent in subject matter, then the procedure can recognize subthemes in a satisfactory manner. If the sentences on the top level form a logical summary of the subject matter of the sample text as a whole, and are not too disjointed, then the procedure can be said to be capable of recognizing main themes.

The matter of subthemes in the sample text from The Genetic Code has been discussed previously; the sentences from the top level of the text structure diagram are presented



as Figure 41. No modification has been made to any of the fifteen sentences, and, from the accessibility of the text in its present form, it is clear that a minor tidy-up to delete connectives ("They realized that . . .", Sentence 2) and to standardize tenses would produce a very readable summary of the thirty-one sentence sample text.

## 4.2 The Corpus of Samples

### 4.2.1 Introduction

The corpus of fifteen samples was selected from across the range of scientific and technical writing, at all levels. It is recognized that fifteen samples do not comprise a conclusive test, but this project, concerned as it was with developing a methodology, was interested primarily in the question "Is it possible to distinguish levels of scientific and technical style using this method of analysis?" rather than the secondary question "Is it feasible?". The author recognizes that for the method of analysis to be proved or disproved conclusively many more samples will have to be analyzed; the author also feels that the basically encouraging results employing the first fifteen samples justify the effort that would be involved in providing an accurate assessment of the feasibility of the method.

The text diagrams produced by PLATEXT for each of the samples occur as Figures C-1 through C-15 in Appendix C. The figures represent both overall successes and overall failures in describing text structure, but even when PLATEXT



was being its most obscure and the text pattern seems to have lost all touch with reality, the analysis will show quite competent insights over the span of a few sentences.

The text diagrams for the fifteen samples can be divided into three categories based on apparent complexity. The first category contains the analyses of texts exhibiting the most complex structures; the second will contain analyses of texts of intermediate complexity; and the third will contain analyses which produced basically linear text diagrams.

#### 4.2.2 Category I: Samples 4, 6, 7, 8, 10

Samples 4 and 10 are from well-known writers who are explaining scientific concepts of natural phenomena to inquiring Everyman (42, 41). Sample 10 has already been discussed in detail. Sample 4 demonstrates how a stylistic device can affect PLATEXT. The first four sentences of the Huxley sample feature parallel construction along the pattern demonstrated by the third sentence: "Only in the water have the molluscs achieved any great advance". The key word in the parallelism is "Only", which is a subordinate relation word in PLATEXT's dictionary.

Sample 7 is from a textbook on genetics (43). The author uses a large number of examples in his exposition; the sample contains two occurrences of "in this case", two occurrences of "i.e.", and one occurrence of "for example" in the twenty-eight sentences of the text. The drop in levels





at the close of the sample text is due to the occurrence of three consecutive sentences each containing one of the aforementioned subordinate indicators.

Sample 8 is another example of this class of writing (44). The notable feature about this sample text from the pages of Scientific American is that it contained the lowest percentage of relation indicators of any sample tested, including the very linear samples of Category III. It also had the longest sentences and highest content word percentage of any of the samples.

Sample 6 is an example of an overly-connected style (30). This extract from an ASIS paper has short-circuited PLATEXT through a much higher than normal percentage of relation indicators.

#### 4.2.3 Category II: Samples 1, 11, 14.

There is considerable room for discussion concerning the dividing line between Categories II and III, although the division between Categories I and II seems reasonably clear. Nor is the population of either Category II or Category III as homogeneous in type as the population of Category I.

Sample 1 is a paper in hydroacoustics which, under slightly different circumstances might find itself in Category III (45). There are four subordinate links (sentences 5-6, 20-21, 30-31, 32-33) due to the occurrence of "since" which the author uses instead of "because".



"Because" is not currently in the dictionary of relation indicators; "since" is employed as a subordinate indicator. Without these four links, the text diagram would be a great deal more linear.

Sample 11 is in reality two samples, being an editorial from Canadian Research and Development, which addresses itself to two distinct topics (46). The break occurs after the sixteenth sentence, and the two topics seem to be discussed in two different styles.

Sample 14 is a short article from Canadian Research and Development (47). A characteristic of short journalistic articles seems to be a dearth of examples. Only the vertical links between Sentences 24-25 and 25-27 are of this type, and "instance" is the relation indicator involved in each case. The remaining vertical links are mostly of the "also" variety, which seems to be another characteristic of these kinds of short informative articles.

#### 4.2.4 Category III: Samples 2, 3, 5, 9, 12, 13, 15

Some of the many samples in this category could well move into Category II with modifications to the dictionary, erasure list, and perhaps to the radius of content matching.

Samples 2 and 3 are a state-of-the-art paper on integrated circuits (48). Both samples are quite linear, with only two vertical links in the entire second sample. The link in the 29-30 case is through the indicator "merely", and the first instance is the sentence "(JK flipflops, for



example)". It is very interesting to note that just as the two halves of the article are close in structural appearance, so are they close in statistics. The average sentence length in the first half is 20.84 words and the content word percentage is 40.78%. In the second half of the article the sentence length is 20.77 words and the content word percentage is 39.89%.

Samples 5 and 9 are examples of "oral" styles (49, 50). Sample 9 is a record of dialogue at a seminar on old English concordances. The characteristics of the text are linear pattern, short sentence length (19.62 words), and low content word percentage (31.14%). Sample 5 is a selection from F. Hoyle's "Science, Society, Action, Reaction", and illustrates an oral style with linear pattern, short sentence length (16.00 words), and low content word percentage (30.73%).

Samples 12 and 13 are taken from newspapers, the latter from the Financial Post and the former from the Sunday supplement to the New York Times (51, 40). The statistics indicate the reasons for the linearity; high content word percentage, low relation word percentage, and, of the latter, very few indicators of the "for example" type. Again, this would seem to be a characteristic of journalism.

Sample 15 is a selection from an IBM manual (52). The characteristics which determine this pattern do not show in the text statistics. The 41.93% content words are of a limited variety, so there is a great deal of content linking. The relation indicators present are mostly of the





coordinate and superordinate type.



## CHAPTER V

### BACKGROUND STUDIES

#### 5.1 Introduction

At the beginning of the project it was evident that the area of concern did not fall exactly in any recognized field of study, and it was not at all readily apparent what discipline contained techniques that would prove helpful. Therefore, a very extensive literature search was necessary. Only two previous studies finally contributed directly to the methodology of the project; they are individually discussed in the Sections 5.2 and 5.3. Other, perhaps better known, studies, techniques, and programs and their relation to this project are discussed in Section 5.4.

As has been seen previously the analytic and synthetic method used in this thesis consists of a procedure employing a dictionary which lists relation-expressing words and their reliability, a list of words which may be regarded as non-informative, and a process of applying the dictionary and the list to a collection of sentences from connected discourse (text) to yield a two-dimensional pattern representative of the intersentence relationships of the text. These intersentence relationships, in turn, are indicative of the formal structure given by the main and subsidiary themes. As stated, the theory of discourse that underlies the procedure and the details of the procedure were developed by the author,



but owe much to discussions with the thesis supervisor and to hints given in two previous papers. The first of these to be discussed is the controversial "Discourse Analysis" by Zellig Harris (53).

This paper has been widely regarded as one of the seminal papers in modern linguistics. But for the purpose of this thesis it is not discussed for the light it may throw on the development of transformational and other modern grammars. It provided the author with suggestions and hints about discourse (connected writing) that contributed to the development of details of methodology and in the following section only those parts of Harris' paper that are relevant are outlined.

## 5.2 Harris' "Discourse Analysis"

In his paper Harris "presents a method for the analysis of connected speech (or writing)". He states that the method requires no semantic knowledge of the text to be analyzed, nor is a listing of the grammatical categories of the words and morphemes necessary, although very helpful. What is required is the distinguishing of different morphemes and words, the junctions between them, and a recognition of boundaries, often but not always punctuation, between the larger units of discourse such as sentences and subsentences. The object of Harris' analysis is the discovery of "patterns of occurrence".

In his further discussion of the first step in the



analysis he mentions the establishment of "equivalence classes" for a sample test. The concept of "equivalence classes" is illustrated by Harris in the following simple example:

The trees turn here about the middle of autumn.

The trees turn here about the end of October.

The first frost comes after the middle of autumn.

We start heating after the end of October.

The first observation Harris makes is that the middle of autumn and the end of October are equivalent because of the "equivalent environments" they enjoy in the first two sentences: The trees turn here about ———. Treating the four sentences as a single connected text, Harris employs the equivalence established in the first two sentences to establish The first frost comes and We start heating as equivalent. By gathering together elements equivalent to each other, Harris then creates "equivalence classes". In the example quoted, the two equivalence classes would be A, consisting of (1) The trees turn here ———; (2) The first frost comes ———; (3) We start heating ———; and B, consisting of (1) about/after the middle of autumn; (2) about/after the end of October. The four sentences are then described as:

$A_1B_1$

$A_1B_2$

$A_2B_1$

$A_3B_2$





where the above symbols correspond exactly to the sentences as presented earlier.

In fact natural discourse is a good deal less amenable to this kind of categorization than the above sample indicates. Harris notes this and describes several accessory techniques which add much sophistication to his original simple procedure and which hopefully will make it possible for an analyst to establish a number of equivalence classes of the type described.

The auxiliary techniques include the treatment of "dependent" pronoun occurrences and of other dependencies related to cross-referencing in sentences. But Harris' use of "dependency" is not the same as that developed for the thesis and will not be discussed in detail here. Harris also discusses the ramifications of using information from outside the text, and introduces the further concept of "grammatical equivalence".

Harris uses all of these techniques in the analysis of the short advertisement in Figure 42a. This figure is helpful in showing relationships between some of Harris' ideas and those found in the thesis. The first steps establish an equivalence class P which includes initially:

Millions

Four out of five people in a nationwide survey  
and another class W which contains:

can't be wrong

say they prefer X--to any hair tonic they've used.



## MILLIONS CAN'T BE WRONG !

Millions of consumer bottles of X— have been sold since its introduction a few years ago. And four out of five people in a nationwide survey say they prefer X— to any hair tonic they've used. Four out of five people in a nationwide survey can't be wrong. You too and your whole family will prefer X— to any hair tonic you've used! Every year we sell more bottles of X— to satisfied customers. You too will be satisfied!

Figure 42a: Sample Advertisement ( after Harris )

PW	Millions of People Can't Be Wrong
BS*I	(the B containing the pseudo-P) Millions of consumer bottles ... have been sold ...
CPW	And four out of five people ... say
PW	they prefer X— ...
PW	Four out of five people ... can't be wrong.
PW	You too will prefer X— ...
PW	your whole family will prefer X— ...
BS*I ( ISB)	Every year we sell more bottles of X—
S*I to P	we sell to consumers
PW	consumers are satisfied
PW	You too will be satisfied !

Figure 42b: Double Array ( after Harris )



The result of the analysis is the "double array" of Figure 42b. Harris admits that ". . . the double array for the advertisement is not interesting in itself . . ." but noted that the interpretation of the contents of the equivalence classes ". . . parallels what one might have said as a semantic critique . . ."; he notes that the uses of the double array of equivalence classes would vary with the nature of the classes themselves, but would doubtless go beyond being a useful adjunct to a semantic critique. Harris also notes the usefulness of his form of discourse analysis in describing some elements of style, as well as the fact that certain logical fallacies in the structure might also be indicated in the matrix of equivalence classes. He also sees another possible use for the matrix as a dictionary of permitted constructions, should one want to add to a text without simultaneously adding to the number of equivalence classes. As Harris remarks, this would be adding to the text without, in one sense, changing its structure. For any investigator of text patterns, it should be noted that the text may be assumed to have a discernible rhythm of its own if a strong pattern is present, and if the equivalence classes are of the same general construction as those furnished by Harris in his examples.

Another possible use Harris finds for discourse analysis is that it ". . . may even show how a particular type of structure can serve new texts or nonlinguistic material." While this writer is not at all confident that





he completely understands Harris' exposition of this last point, the impression is conveyed that Harris feels there may be some optimum structure for the expression of certain material, textual or otherwise, and that discourse analysis of this type may help find that structure.

### 5.3 Relation to PLATEXT

The content-linking component of PLATEXT is related to, partially suggested by, and partially developed from Harris' approach to "Discourse Analysis". It seems that Harris' procedure is very close to a philosophical logical analysis of the Carnap or Tarski variety in that the analysis extends internally to fragments of statements, to phrases, and even to individual nouns and pronouns. Harris' analysis is not of the meaning of the forms being investigated, but rather establishes equivalence classes based on function within the text with the aid of specified grammatical considerations. The classes are symbolized without regard to sentence boundaries, and the "discourse analysis" consists of seeking repetitive patterns among the strings of class symbols. The repetitive sequence may exhibit itself as a sentence, or less than a sentence, or more than a sentence, or all of these in various parts of the text. In Figure 42b, the repetitive sequence and basic pattern is "PW", and a primitive grammar was employed in the establishment of the equivalence classes of that particular example. Whether Harris saw, at this point, such a grammar as an integral



part of all discourse analysis is not clear.

In contrast, the pattern sought by PLATEXT is not the repetition of (classes of) phrases, but the picture of how localized sets of content words are ranked relative to each other, with respect to dependency, or importance, or relevance to the main theme. Because the set boundaries in the analysis are identical to the sentence boundaries in the text, the populations of the sets are author-defined. PLATEXT does not take into account the internal ordering of the elements of the sets, and the elements (content words) of any given set may contain several examples of the same content word.

PLATEXT uses no grammatical information other than the separation of content words from text by the employment of an exclusion list, suitable for all Standard English, with some reservations discussed in a later section. Except for this type of selection, the content matching in PLATEXT is "blind". Even so, a reasonable alternative to Harris' methods and effects is achieved. For example, in the procedures of PLATEXT a long phrase containing four content words weighs twice as heavily in establishing a content link as a weaker phrase containing only two content words.

The idea of taking localized sets of content words, comparing them to other sets of content words found in some fixed neighbourhood of discourse, and drawing a link between the two sets with the most elements in common, seems logical



and appropriate, particularly when the criteria for set boundaries are author-imposed.

In sum, the debt to Harris' "Discourse Analysis" is expressed by the horizontal links in the text structure diagrams produced by PLATEXT.

#### 5.4 Jacobson's Sentence-Connecting Routine

The second important influence from the literature on the development of the theory discussed in Chapter II of this thesis is a paper by the late S. N. Jacobson "A Modifiable Routine for Connecting Sentences of English Text" (54). In the previous section it was noted that Harris' paper has been sometimes regarded as obscure by those using it to trace the development of various modern grammars. Similarly, Jacobson's paper presents certain difficulties in interpretation. It is not always consistent, and no further work which might aid in clearing inconsistencies developed from it.

Jacobson's paper uses the concept of relative dependence of sentences in text, and describes text structure in graphical form. Jacobson finds evidence of the structuring of sentences in connected discourse in the varying degrees of dependence exhibited by these sentences. He devotes part of the paper to a discussion of sentence outlines and of the manner in which they demonstrate the relative dependence of the constituent sentences of a text. Sets of examples (one such set is included here as Figure 43)



- 1) However, this is merely a psychological theory.
- 2) This is merely a psychological theory.
- 3) This is a psychological theory.
- 4) Gestalt theory is a psychological theory.

Figure 43: Sentence Dependency Hierarchy ( after Jacobson )





are given to demonstrate this gradation.

In the set of sentences (Figure 43) the words and phrases that determine each sentence's text-dependency are examples of what Jacobson calls "clues". In addition to marking the containing sentences as more or less text-dependent, the clues may indicate the sentence in the text to which the sentence in question is related. Jacobson's sentence-connecting routine relies on the properties of these clues. His definition is as follows:

in general, a potential clue is either

- 1) A word or construction which can be replaced by any one of a small list of items of which it is a member, such that this replacement does not modify the sentence structure in a significant way. An example would be the word this which can be replaced by the word that in the first sentence of (Figure 43).
- 2) A word or construction which can be deleted without destroying the grammatical function of at least one of the constructions in which it occurs. Examples would be the words however and merely in (Figure 43). Their deletion still leaves a grammatical sentence.

Clues relate a dependent sentence to the other sentences in the text in two ways: (1) they arouse the expectation that there is another sentence in the text to which a dependent sentence is especially related, and (2) they permit the reader to make a prediction about this other sentence so he can recognize it if he encounters it in the text.

Jacobson uses these predictions to establish links among the sentences of the text. There are two types of prediction in his routine. The first seeks the reoccurrence of a noun that has occurred in a noun phrase marked as a clue. Jacobson offers as an example the phrase this task, which, if it occurs, suggests that the noun task has occurred earlier in the text. The second type of prediction seeks the



reoccurrence of a clause type. As an example Jacobson lists clauses beginning with the clue however. These searches are much less particular, seeking not a specific clause but merely an example of a specified type.

The immediate task of the clues is to establish what Jacobson terms the "principal path" through the text. These are the most important sentences in the text; Jacobson states these sentences are those an efficient auto-abstracting scheme would have selected as being representative of the text. In terms of a sample sentence outline of the second chapter (Figure 3) the principal path would be the string of sentences labelled A, B, C, . . .

Jacobson refers to his method as "computational", and, while it is not clear from his paper to what extent the process has been implemented on a computer, he makes it plain that the sentence-connecting routine was designed with that in mind from the outset. As noted, no follow-up references were found.

Jacobson's system consists of four parts:

- 1) dictionary of clues. As described above, "clues" make predictions, and list conditions under which the prediction is satisfied.

- 2) routing procedure. (See Figures 44, 45, and 46). The "routing procedure" directs the search for predicted entities, and assigns links between the sentences if the prediction is satisfied. The cluing and linking procedure may involve modifying established links.



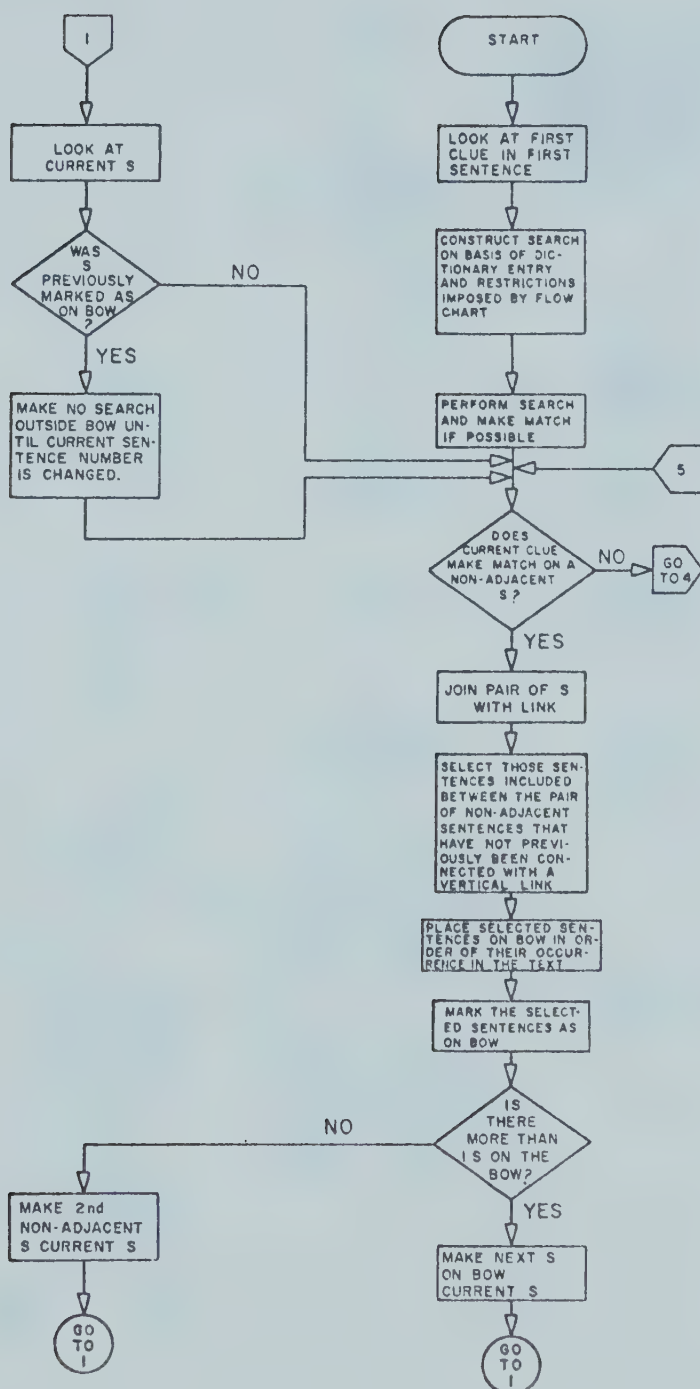


Figure 44: Routing Procedure I ( after Jacobson )





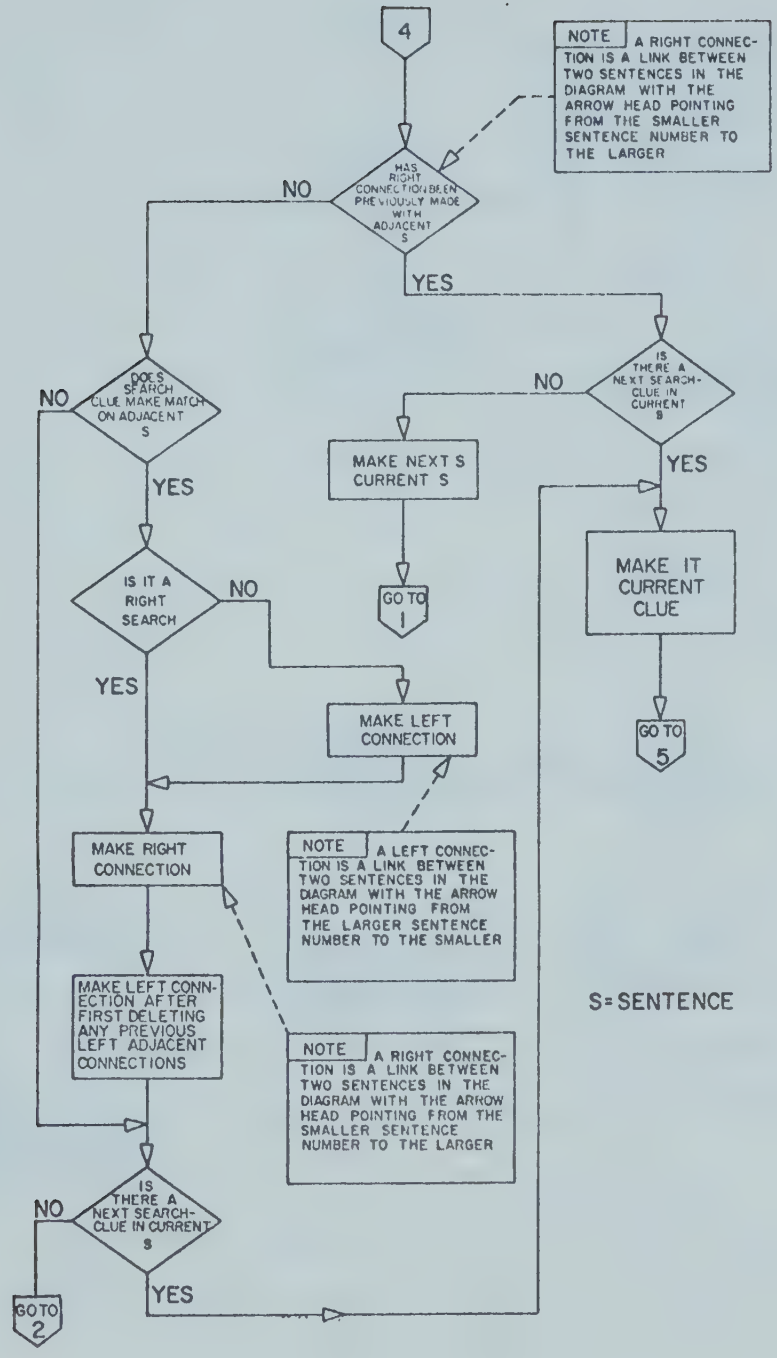


Figure 45: Routing Procedure II ( after Jacobson )



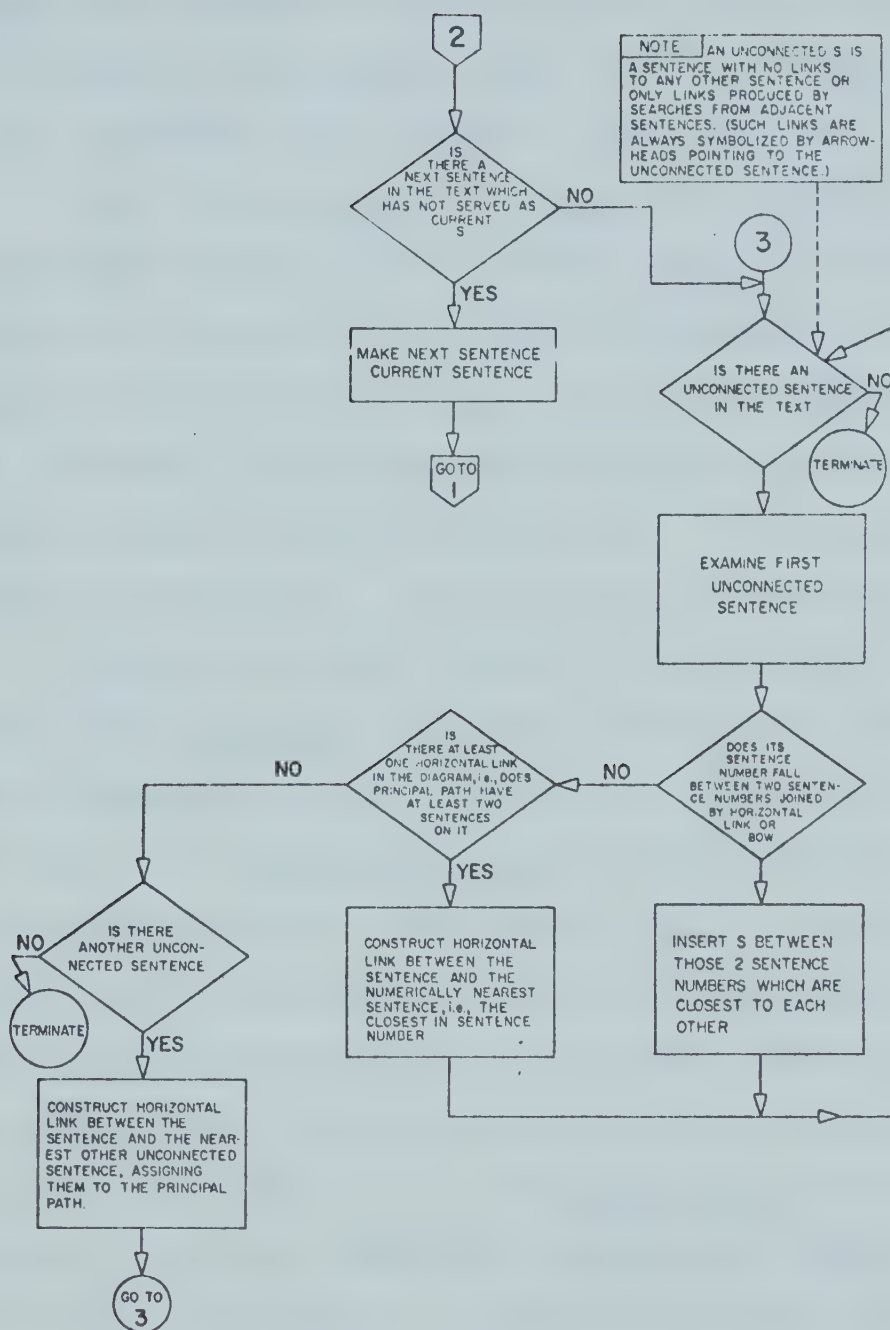


Figure 46: Routing Procedure III ( after Jacobson )



3) sample text. The sample text undergoes a syntactic analysis, but Jacobson's paper does not indicate whether this is manual or automated, only that it is necessary.

4) a procedure for creating a "text diagram".

The "text diagram" is Jacobson's graphical version of the sentence outline and is a result which is of much interest to this thesis. The basic symbols of the text diagram are described in Figure 47, the principal path is shown in Figure 48, and the text diagram for a twenty-eight sentence sample is shown in Figure 49. The sample text is included as Appendix E, for purposes of comparison.

The basic symbols (Figure 47) can be explained as follows: The arrowhead indicates whether the issuing sentence exhibits dependence upon a sentence to come or a sentence since past. The horizontal link is used to connect sentences on the principal path. The vertical link is used by Jacobson to indicate the intersentence relations corresponding to the sentence outline sequence 2a, 2b, 2c, and also to indicate the intersentence structure corresponding to the sentence outline sequence II, 2, 2a. The dotted link is used by Jacobson to join two apparently unconnected sentences to preserve the integrity of the structure of the principal path. The bow and bow-node are representations of digressions in the text, which are treated as independent subtexts. Bows and bow-nodes will be discussed in the next section.

Figure 48, the "reduced set of paths through the text

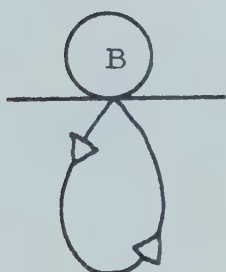




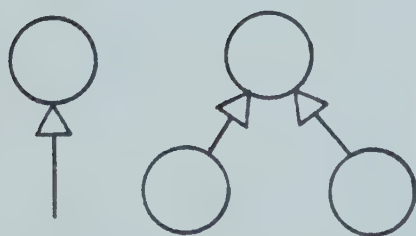
Arrowhead indicates direction of dependence between sentences in diagram.



Horizontal link between nonadjacent sentences which are not included between nonadjacent sentences.



Bow and B-node indicate links between sentences included between nonadjacent sentences. B-node is the dummy sentence thru which these sentences are joined to the principal path.



Vertical links between adjacent sentences. Note that a vertical link may be drawn on a slant.



Dotted link(s) join(s) an otherwise unconnected S to the text paths so it is ultimately linked to the principal path.

Figure 47: Symbols ( after Jacobson )







Figure 48: Principal Path  
( after Jacobson )



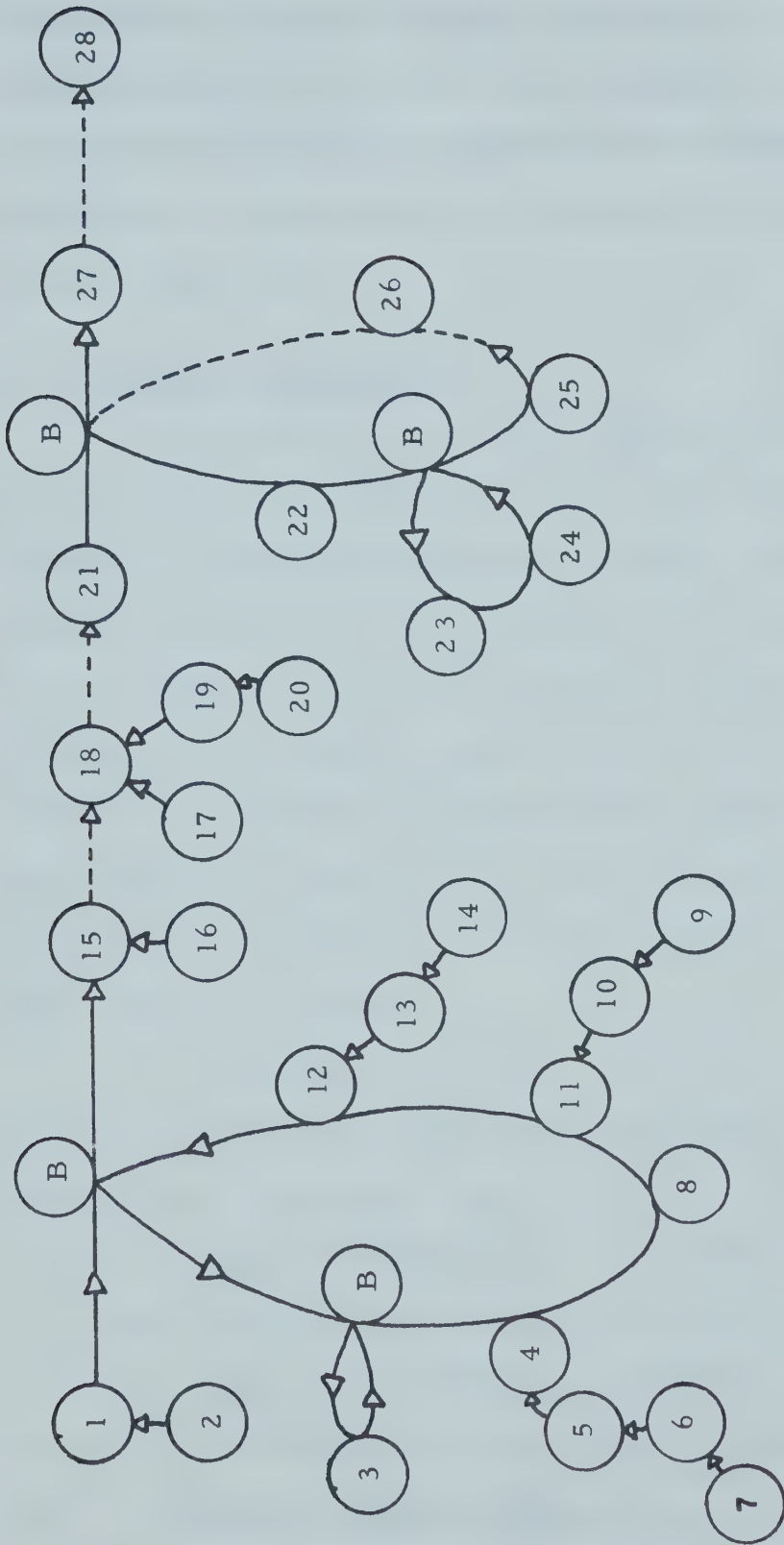


Figure 49: Full Text Diagram  
( after Jacobson )



(principal path)", is the string of sentences from the sample text judged to be of primary importance. Figure 49, the "text diagram" or the "full set of paths through the text", is the pictorial representation of the twenty-eight sentence sample text, as produced by Jacobson's routine for connecting related sentences.

## 5.5 Relation to PLATEXT

It is obvious that the theory of discourse developed in this thesis is indebted to the ideas expressed in Jacobson's sentence-connecting routine. However, there are certain very specific differences in the overall theories and in some aspects they are not closely related. The precise relation of the thesis theory and of its implementation in PLATEXT to Jacobson's "computational approach" will now be discussed with respect to 1) the dictionary of "clues", 2) the routing procedure, 3) the text sample itself, and 4) the resultant text diagram.

1) dictionary of clues. Jacobson's dictionary of clues is basically a content-linking device rather than a dependency or structure-seeking device, despite the implications of his early examples. The clues are "predictions" (in effect, tests) concerning either nouns and noun phrases, or clause types. For example, encountering this task sends the routine to seek an earlier use of task, and the occurrence of However, \*clause type A\* . . . sends the routine to find a previous occurrence of \*clause type A\* in the text.





Without exception, the linking arising from successful predictions on the principal path is horizontal.

In contrast, the dictionary in PLATEXT contains clues, or dependency relation words, but no tests. The mere occurrence of one or more of these words in a sentence is considered, to a greater or lesser degree, significant. This method, while much closer intuitively to the development suggested by Jacobson's early examples, might be intrinsically more inaccurate than a predict-and-search method. In practice employing many dictionary clues of various weights and taking an average for each sentence has proved at least as accurate as predict-and-search. The dictionary in PLATEXT establishes vertical linking on all levels, and thus contributes to a more satisfactory structure.

2) routing procedure. The basic function of Jacobson's routing procedure is the management of the predict-and-search mechanism. Each dictionary entry contains full information on conditions for satisfying its prediction, plus the range the search is to cover, plus the direction of the search (forward, backward, or both). At any given time in the text processing, information on all current links must be available. A link (satisfied prediction) between Sentence  $N$  and Sentence  $N + K$  will be negated by a discovered link between Sentence  $N$  and Sentence  $N + K + P$ . The link between  $N$  and  $N + K + P$  isolates the  $K + P - 1$  included sentences; no links are allowed between them and any sentences outside the interval  $N$  to  $N + K + P$ . Any links



discovered internally among Sentences  $N + 1$  to Sentence  $N + K + P - 1$  serve to organize the interval into bows and subbows (see Figure 49). This seems to indicate that Jacobson's model of discourse is less satisfactory or less fully worked out.

The interpretation of the significance of an established link and the consequent loss of intervening links is not carried over into PLATEXT. All links in PLATEXT are considered in the final judgement of the predominant nature of the sentence. A more fundamental difference between Jacobson's routine and PLATEXT is found in the separation of the roles of the horizontal and vertical links. The application of these defined roles is universal in PLATEXT; fragments of text beneath the top level (principal path) are structured according to the same rules as those on the main theme. The view of overall text structure has been kept consistent.

3) text sample. The text being fed to Jacobson's routine has obviously been pre-edited; specifically, the identification of noun phrases and main verbs has been made. No mention is made of computerization of this phase of the system, or of the possibility of its computerization. Considering the complexity of the task, as it is indicated by careful consideration of the article, automation seems remote at best. It certainly could not be automated so that the required text pre-editing could be done simply and economically.



4) text diagram. Jacobson's text diagram places tremendous significance on a small number of satisfied predictions, at least in the presented example. Recalling Jacobson's definitions of horizontal, vertical, dotted, and bow linkages, one realizes that the text diagram of Figure 49 is based on two satisfied predictions (isolated in Figure 48, the "principal path"). The only linkages on the principal path are 1 to 15 and 21 to 27, and, as described above, these links define to a great extent the rest of the text structure. Again, PLATEXT's system of summing many links would seem to be preferable to Jacobson's technique.

The most striking feature of Jacobson's diagram is that the further one moves below the principal path, the less the diagram represents a sentence outline in any interpretation. In fact, Jacobson's implementation seems only tenuously connected to his original conception. Again, a more accurate assessment of the effectiveness of his routine would have been possible had Jacobson chosen an example of expository writing, rather than of creative writing.

To sum up, the significant debt to Jacobson's "A Modifiable Routine for Connecting Sentences of English Text" is expressed by the vertical links in the text structure diagrams of this thesis.

## 5.6 Related Fields of Study

### 5.6.1 Introduction

The literature search conducted at the outset of this





project resulted in the investigation of a large number of procedures that might have been considered contributors to the methodology of the thesis. Although these investigations were disappointing and offered little to the thesis, the most significant of them, whether they consist of procedures, programs, algorithms, or distributions are briefly discussed in this section to furnish some additional background; the relevance of each to the methodology and objective of PLATEXT is outlined.

#### 5.6.2 Text Handling Methods Based on Statistics

This discussion will divide the cited literature covering statistical approaches to text handling into two groups. The first group contains papers that deal primarily with statistical distributions, usually of word frequency classes. The second group consists of approaches based on statistical inference, usually applied to specific samples of text relevant to an individual authorship problem.

5.6.2.1 Bibliometrics. The term "bibliometrics" was revived by Alan Pritchard (55) and popularized by Robert Fairthorne in 1969 (56), who paraphrased it to denote the "quantitative treatment of the properties of recorded discourse and behaviour appertaining to it". The study of statistics of large corpi of text has enjoyed much popularity in the last two decades; a listing of even a fraction of the proposed distributions and their relation to word frequency classes would be inappropriate here. What is important to





note is the increased attention paid to these specialized mathematical functions by workers in information retrieval, particularly those concerned with indexing and storage. The search for methods of characterizing the distribution of words in text is further justified now by the number of words in text which are to be found in mechanized systems. Any information on the distribution of words that can be reliably incorporated into indexing and storage systems and that can lead to the saving of even a small percentage of effort in time or storage space can effect considerable economies. Bibliometrics pays particular attention to masses of text which fail to exhibit the predicted distributions; a consistent non-conformity of a particular range of samples could well turn out to be enlightening and possibly immediately useful. It is this type of non-conformity in distributions that interested researchers seeking methods useful for isolating particular classes of documents.

A few of the important distributions are outlined in the following sections.

Zipf's Law. The prototypical bibliometric distribution was popularized by George Kingsley Zipf in 1949 (57) after being introduced by J. B. Estoup in 1916 (58). The "Law" holds the relationship

$$r * f = C$$

to be true, where  $r$  is the rank of a particular word in a listing of distinctive words arranged according to decreasing frequency in a given text, where  $f$  is the frequency associated



with that word, and  $C$  is a constant for that particular text. Zipf himself claimed a universal value of  $0.1 \times \text{text length}$  in words for the constant as applied to English text. In his book Human Behavior and the Law of Least Effort (57) Zipf offered a psychological-sociological explanation for the regularity in lexical statistics he found, and this explanation seems to have confused the issue, perhaps because its subjective nature did not admit the rigorous testing of its validity.

The most commonly encountered form of Zipf's Law is

$$f(r) = k/r$$

where the rank of a word,  $r$ , is the number of words, including itself, that have at least the same frequency,  $f(r)$ , of occurrence, and where  $k$  equals (approximately)  $1/10$ .

Mandelbrot. The most prolific champion of the Zipf-type distributions has been Benoit Mandelbrot, who has offered various mathematical models of discourse which lead to relationships of that family (59, 60, 61). The models are variants of two basic types.

The first, subsequently developed by George Miller (62), assumed text to be generated in a letter-by-letter random fashion with a fixed probability of occurrence for each of the letters and the space. This defines a Markov chain whose states are in the different letters of the alphabet, plus the space. The results of allowing the model to proceed in its prescribed manner are of a type that can be accurately described by Zipf's Law. In his



second model, Mandelbrot de-emphasizes "rank" and proposes instead "cost", one measure of which he declares to be word length. The parameter "entropy", associated by Mandelbrot with the distribution of word probabilities, is maximized and the generated text put in its "most probable" state. The resulting relationship can be expressed as

$$f(r) = k(r + c)^{-\theta}$$

which can be adjusted to a closer fit of the data than the Zipf distribution.

Simon. Another class of functions for the description of lexical class frequencies was proposed by Herbert A. Simon in a series of papers over a five year period (1955-1960) (63, 64, 65). Unlike the Zipf-Mandelbrot model which generated text letter-by-letter, the model proposed by Simon generates text word-by-word. The distribution advanced by Simon met academic opposition in Mandelbrot, and a lively debate on the relative merits of the two distributions graced the pages of the journal Information and Control (60, 61, 64, 65) over a period of time. That two so widely diverse models and functions could more or less plausibly claim to account accurately for the identical distribution of frequency classes is interesting.

Herdan. The fourth family of distributions in this field is that proposed by Gustav Herdan (66, 67), in opposition to the models of Zipf-Mandelbrot. The most viable of Herdan's proposals would seem to be the Waring-Herdan distribution, which receives its best discussion in "Lexical







Distribution Reconsidered: the Waring-Herdan Formula" by Charles Muller (68). Muller applied this distribution, which is unrelated to the previous three families, to samples of text; again, excellent fits of data to predicted distribution were reported.

The field of bibliometrics has been covered by several excellent reviews and bibliographies, the review by Mandelbrot (69) and Fairthorne's survey (56) being particularly important.

Determining the significance of all these distributions provides fertile ground for discussion. The surprising agreement of a surprising variety of distributions should, intuitively, provide some statistical insight into facets of information handling. Further interesting research and results in this area will undoubtedly emerge in the future. But, at this point, statistics in the mass have not provided a methodology useful in achieving the objectives of this thesis.

5.6.2.2 Statistical Inference. The application of statistical decision theory to the problems of "comparative" style analysis of individual author's works has proved to be very fruitful. The fundamental problem in such applications is always the selection of test parameters which will be truly indicative of a significant difference in the candidate authors' modes of expression. Typically, many potential statistical measures are dependent on sample



length, the choice of which is not always left to the investigator. Two examples of significant stylostatistical studies are outlined below to illustrate the basic nature of the decision process as applied to questions of disputed authorship.

The first example is the study of F. Mosteller and D. L. Wallace of the authorship of the disputed Federalist papers (70). The available samples of the candidates' style were short and subject-oriented. Consequently, previously successful methods based on noun-comparisons could not be employed. Further, the mean sentence lengths exhibited by known samples of the writing styles of the candidates were not significantly different. Mosteller and Wallace eventually established a set of function words that one candidate favoured, and used these as a basis for the statistical test.

An extension of the Mosteller-Wallace technique was used by A. Ellegard in his investigation of the Junius letters (71). Ellegard took known examples of his principal candidate's writings and compared his usage of hundreds of test words to the frequencies of usage exhibited by his contemporaries. Ellgard took the results of these comparisons and applied them to the frequencies of occurrence of test words in the Junius letters and found strong agreement.

The two examples cited demonstrate some of the processes of stylostatistics, and the type of problem which can be frequently solved through their techniques. In



general, it can be said that the applicability of stylo-statistics varies with the specificity of the problem. Stylostatistic methods which have been developed to magnify differences between authors writing on the same subject are not as a rule effective if the authors are writing about different subjects, or on a different level for a different audience. And, while the computer has made extensive comparisons easier, it has not eliminated the preliminary manual work that has to be done to establish the appropriate statistical parameters. In the case of the objectives of the thesis, the required stylostatistic would ideally be sensitive to "levels of style" (as described in Chapter 1) reflecting the differing intended audiences; it would remain insensitive to the free substitution of authors and subjects in view of the extremely wide range of samples likely to be encountered under "science and technology". It should be noted that "readability" parameters accomplish this with respect to the literate versus illiterate audience dichotomy, but these parameters (mean word length, mean sentence length, etc.) cannot differentiate levels of style in the context of the investigation of this thesis.

### 5.6.3 Content Analysis

Several general content analytic programs were investigated at the outset of this project, and were found to be in some senses inadequate and in others overadequate for the purposes in mind. The principal programs involved





were MAPTEXT, VIA, and GENERAL INQUIRER.

The first two are programs developed by Dr. Sally Sedelow in collaboration with D. G. Bobrow and T. L. Ruggles respectively (72, 73). The earlier program, MAPTEXT, was written to provide the literary researcher with a means of reducing text to simple graphs. The program illustrates graphically the positional relation between specified textual elements (such as punctuation, function words, keywords, etc.) by printing distinctive symbols at each point in the text where an instance of that element is found. The result of the analysis depends, of course, on the particular elements with which the researcher was working, and what he or she was doing with them, but in any case the regular or irregular sequence of elements in the text will be converted by MAPTEXT into a readable pattern. And, if the author of the text has any distinctive trait involving the distribution of punctuation, or function words, or pet phrases and clichés, the researcher can at once isolate it, and produce "hard copy" evidence (See Figure 50).

VIA (for Verbally-Indexed Associations) is a far more complex program based on word associations. The program, which incorporates a version of MAPTEXT, was developed to aid in the search for patterns that would indicate themes recurrent in the text. The patterns that VIA keys on are the reoccurrence of words and phrases in loose proximity. The program begins by going through the offered text and noting, in index form, the frequency of occurrence of each





Praeger Chapter One

1 ---MS---D---M---M---D-MS..  
2 ---A-W---5---M---M---W---0---  
3 -G---G---W---G---W---  
4 ---M---M---D-U---W---  
5 ---  
6 ---M3---S---S---G---M---  
7 ---W---  
8 ---D-MS---W---G3---  
9 ---A---D---G---0-M---  
10 ---M---M---G---MS---W---M-S---M-..  
11 ---MS---M-G---M---G---M---  
12 ---M---  
13 ---W---W---W..  
14 ---D---D---A---D---M---W..  
15 ---M---W---S---M---S---M---  
16 D-M---M---4M---D---MS---MS---..

Figure 50: MAPTEXT ( after Sedelow )



word. Content words of high frequency are matched against each other to test for co-occurrence and the results of this, together with the results of a similar check on related words the machine has saved from previous analyses, are printed out in the form of a "cross-referenced thesaurus". In method and intent, VIA is much closer to the GENERAL INQUIRER, a content analysis program, than to MAPTEXT, a stylistic analysis program. Figure 51 is an example of output from VIA.

The GENERAL INQUIRER is the name given by its makers to a commercially available content analysis program. Developed by R. J. Stone, D. C. Dunphy, D. M. Ogilvie, et al. the GENERAL INQUIRER depends on a manually prepared dictionary in which the entries are associated with one or more "categories", which number usually between fifty and one hundred (74, 75). For example, in

To be, or not to be,--that is the question:--  
 Whether 'tis nobler in the mind to suffer  
 The slings and arrows of outrageous fortune,  
 Or to take arms against a sea of troubles,  
 And by opposing end them?

the words "or" (both of them), "question", "whether", and the punctuation mark "?" might be tagged, i.e. put into the category named, INDECISION. Similarly, "slings", "arrows", "take arms against", and "opposing" might be tagged VIOLENCE, and "suffer", "outrageous fortune", and "sea of troubles" might be tagged DISQUIETUDE. Clearly the effectiveness of



298	DRINK		11
		-----CUP	
		-----LIQUOR	
		-----POTION	
	DRINKS		
312	* EARTH		9
	*	HEAVEN	5
		EARTH	
		STARS	
		LAND	
	*	NATURE	5
		LIFE	
		UNNATURAL	
		WORLD	

A Sample of VIA'S Output for Hamlet

Figure 51: VIA ( after Sedelow )





the program depends largely on sagacity in the choosing of categories and their members. The initial defining of categories is essentially a manual procedure.

The output from the GENERAL INQUIRER can be varied according to the needs and preference of the user. A typical request might be for a copy of the text overlaid with a corresponding set of patterns of category reoccurrences, which would roughly correspond to a MAPTEXT and VIA combined. From the point of view taken in this thesis, the disadvantages of the GENERAL INQUIRER system are two-fold. The first is the reliance on a manually prepared dictionary. Unless one is fortunate enough to be working in an academic area for which someone has already constructed a GENERAL INQUIRER dictionary, one is faced with a singularly tedious, but absolutely critical and very subjective, task. The second disadvantage is that in trying to be universally applicable, GENERAL INQUIRER has become too large and complicated for the typical application. The complexity of the program is a consequence of the generality of the program. In support of this approach it should be noted that the GENERAL INQUIRER has been used for analyses of everything from inaugural addresses to collections of suicide notes.

The disqualifying factor for the content analytic programs as regards this thesis project is the same factor which disqualifies stylostatistic methods. As noted in the previous section, this project demands in effect, content independence. The intended field of application being



"science and technology", dictionaries containing categories of content words as required by VIA and the GENERAL INQUIRER are patently impossible to construct. This does not suggest that content analytic techniques are not relevant to this study, or that PLATEXT's methodology could not be used to augment existing content analytic techniques. The potential cross-fertilization is an area that should be further investigated, and will be discussed briefly in the Concluding Discussion.

#### 5.6.4 Auto Abstracting

The relationship of auto-abstracts to the thesis project has already been mentioned. The first investigator to propose that a computer could produce usable abstracts of documents was the late Hans Peter Luhn. In his original paper, the process proposed selecting sentences from the document by ranking all the non-function words in the text to arrive at the words most indicative of the content of the document, then selecting the sentences which contain the greatest occurrence of these words (76).

The advances in automatic abstracting since 1958 have resulted either from attempts to overcome some of the shortcomings of the basic Luhn method, or from attempts to utilize a process fundamentally different from Luhn's statistical approach. Each involves an increase in sophistication and complexity, with a resultant rise in time and cost per abstract. These experiments, however, are



independently valuable for their relevance to problems in the analysis of literary structure. S. N. Jacobson's paper, for example, develops his sentence-connecting routine in the context of an automatic abstracting procedure. Other interesting work, even if not closely related to the thesis, has been carried out at Ohio State by Petrarca and Rush (77, 78, 79, 80).

It is evident that PLATEXT is quite suited to the production of automatic abstracts, although such was not the intention of the program. Briefly, the abstract would be formed by the sentences occurring on the top level of the sentence diagram. Possible modification to the program making it more effective in this specialized application will be suggested in the Concluding Discussion.

#### 5.6.5 General References

The project required, as well, background reading of a more general nature than the specific studies referenced to this point. Some of the more invaluable general works are included in the bibliography as Nos. (81) through (85).





## CHAPTER VI

### CONCLUDING DISCUSSION

#### 6.1 General Comments and Immediate Relevance

The objective of this thesis was the development of a methodology for the recognition of different levels of scientific and technical literature. In fulfilment of the objective, a technique for recognition and description has been devised. Its refinement would go far towards 1) further defining and delimiting the problem, and 2) as specific situations involving this type of precision difficulty emerge, solving the problem for those situations.

The samples processed by PLATEXT have demonstrated the soundness of the basic procedure, and have suggested paths for further research and development that might be followed to attain the long-range network-oriented goals outlined in Chapter 1. A serendipitous outcome, in view of the original thesis objective, is the immediate relevance of PLATEXT to automatic abstracting, content analysis, and other types of stylistic studies. In the following sections, each will be briefly considered.

##### 6.1.1 Automatic Abstracting

As mentioned earlier, the sentences on the top level of the text structure diagram are assumed to be the most important in the analyzed text sample. These sentences





could form an abstract or summary of the text, and, if the application of PLATEXT were to be solely the production of auto-abstracts, a number of changes to improve the program's effectiveness might be considered.

The first modification would doubtlessly be the development of a more specialized dictionary, with emphasis on the superordinate relation words. Because the specification of levels below the topmost is deemphasized, the subordinate and coordinate dependency indicators could be reexamined for their relationships to possibly independent sentences around them.

A second modification for this special application might be the respecification of some of the text structuring rules. For example, PLATEXT now places sentences which contain no dependency indicators or content links on the top level; in certain varieties of discourse (e.g. dialogue) this rule accounts for most of the population on the top level. For auto-abstracting, this rule might profitably be waived for at least those kinds of discourse which, under the analysis of PLATEXT, exhibit a paucity of content links.

Another area for experimentation might be the radius of content linking in PLATEXT. An increase in the radius of content matching would decrease the possibility that a sentence could exhibit no content links; under the circumstances of an increased radius of matching, the second modification suggested above might not be required.

It should be noted, particularly with respect to



this application, that with an appropriate dictionary and erasure list, PLATEXT is multilingual. The content-linking component of the program, being "blind", matches both within and among physics, Basque, and gibberish with equanimity.

In general, the formation of an auto-abstract is a subset of PLATEXT's operations on samples of discourse. The further development of this particular facility would require a great deal of testing of samples from the field to provide data for empirical adjustments to the program, and a corresponding reduction in the generality of the program.

#### 6.1.2 Content Analysis

The relevance of PLATEXT to the typical problems in the field of content analysis is readily apparent. As can be seen from Figure 51, content analysis programs can isolate a theme, as expressed by any of the words, phrases, or punctuation specified in the dictionary category for that thematic collection, and can document the co-occurrence of two or more theme-words within a specified distance of one another. When used in conjunction with a content analysis program such as VIA or GENERAL INQUIRER, PLATEXT could therefore immediately provide a graphic description of the relationships between the themes in terms of dominance and subdominance.

A more complex system incorporating the general principles of both PLATEXT and the content analysis programs would have usefulness far beyond the usual bounds of content analysis, and will be further, but indirectly, discussed in



## Section 6.3.

### 6.1.3 Style Analysis

The definition of the problem to which this thesis is addressed, the development of a theory of discourse, and the implementation of a model of that theory all fall within the field of style analysis, and the resultant computer program has a surprisingly general applicability throughout the field. In its present form, PLATEXT can serve as an aid to the teaching of technical writing, and its application to the universe of non-technical writing has not been even tentatively explored. There is no reason not to expect PLATEXT to be at least as relevant as some of the content analysis programs, and PLATEXT is in addition much easier to use. Again, the application of PLATEXT to other specific problems within the field of style analysis might justify the modification of the program to make it more effective in dealing with one or another of these problems. The modification required might include revising the dictionary of relation words and the erasure list, particularly if the period of English literature under consideration is not contemporary. Given the richness of form and content of English literature at each stage in the evolution of the language, a tool for stylistic analysis based on a new approach should find beneficial employment for a considerable time. As with automatic abstracting, one of the important features of PLATEXT is the ease with which it can be adapted







to foreign, dead, or obscure languages and dialects.

## 6.2 Immediate Development

PLATEXT has to this point in its evolution remained a program of reasonably universal applicability, at the expense of some measure of effectiveness when applied to specific problems. Further modifications will doubtless be in the direction of specialization in one field or another, but some general observations can be made.

The next steps in the development of PLATEXT should include a revising of the erasure list, and an augmenting of the dictionary of relation-indicating words. As mentioned in Chapter 3, the relation list was taken from Kucera and Francis' A Computational Analysis of Present Day American English; the salient points are that the list is Present Day and American. The million words from which this Brown University list was compiled were uniformly written in 1961, and, while one intuitively expects that a million words of 1972 American English collected in the same manner would produce about the same five hundred most-frequent words, the prospect of American English in 1984 does not inspire the same confidence. Mention has already been made of the potential problems in applying PLATEXT in its present form to period pieces of English literature, which have no responsibility to 1961 frequency distribution curves. To this writer's knowledge, only the broadest kind of statements can be made about the way our language is changing from day to



day, even within the confines of an area such as scientific and technical writing.

The problem of obtaining a meaningful erasure list is compounded because our environment is Canadian. Words such as: STATE, STATES, AMERICAN, UNITED, PRESIDENT, WASHINGTON, etc. find themselves, not surprisingly, in the upper echelons of frequency classes in the Brown University list. The ranking accorded these words simply does not apply in the Canadian environment; presumably we have words of our own to occupy these positions. As of this writing, however, no one knows what they are, for no Computational Analysis of Present Day Canadian English exists, although a dictionary of Canadian English has been compiled (86) and had provided a basis for some computer investigation (87).

The dictionary of relation-indicating words can be made more effective for each specific application; it might also be made more effective by increasing its size through a thorough further manual analysis of scientific and technical writing. The inclusion of more relation-indicating words would place additional strains on the system of weights that reflect the reliability of the indicators; in response, the weighting scheme would have to become more sophisticated. The obvious relation words, as included in PLATEXT's dictionary, were relatively easy to select from text, but a complete list of relation words and their associated weighting factors will be much more difficult to compile. Presumably the dictionary of relation words would be carried to completeness



only with reference to specific fields or problems.

### 6.3 Two Steps Past PLATEXT

This thesis has not been concerned with the psychological and psycholinguistic aspects of the theory of discourse outlined in Chapter 2. But these two sciences might address themselves to basic questions suggested by the theory: questions such as why do engineers tolerate, even prefer, a linear structure characteristic of a "dry" style for their own education within that field, when the same information and possibly more could be transmitted in the more complex structures characteristic of "popular" science (Scientific American, Isaac Asimov, Willie Ley, etc.)? Conversely, why do people "just interested" in a subject prefer popular science writing to the appropriate textbook on the subject?

To answer the above questions, one must first confront the problem of describing the fundamental differences that exist in styles, and their effect. The problem reduces to an examination of what information is transmitted by the sentences of each. In terms of a simple dichotomy, "The box is in front of the table" must be contrasted with "Unfortunately, the box is in front of the table". To the average robot, the second statement does not contain any more information about the environment than does the first statement, yet our intuition, and perhaps common sense, tells us that the second statement is the bearer of more information





than the first. However, neither intuition nor the literature can tell us how to measure the amount of extra information borne by the second statement.

The difficulty is that the extra information supplied by the second statement is very subjective in nature; it cannot be separated entirely from the speaker's or writer's frame of reference, a personal property. In a word, the information is connotative.

The importance of this kind of information cannot be overemphasized. Most of the world's political and social decisions are made on the basis of connotative information, yet techniques for handling this sort of information have been neither described, nor, as far as is known to this writer, have they been specifically sought. Two kinds of programs are capable of a degree of connotative analysis, one being PLATEXT and the other being the larger content analysis programs such as VIA and GENERAL INQUIRER. Such content analysis programs have so far been limited to elaborate tagging of words and phrases in text and some of the dictionary categories employed in this tagging have sought "connotative" words. PLATEXT, in its representation of levels of relevance, uses a two dimensional diagram to convey content and one connotative facet, i.e. the degree of emphasis associated with each sentence.

PLATEXT and the typical content analysis program can be compared and contrasted in the following manner: PLATEXT contains a dictionary which tags three categories





of words; the content analysis program contains a dictionary of many categories, each of which may tag many words. PLATEXT acts on sentences as units, and all dictionary entries from all three categories are considered when deciding the fate of the sentence. As a rule, content analysis works on no such fixed unit, although when analyzing extremely structured forms such as poetry, a line-by-line analysis becomes meaningful. In PLATEXT, content linking is done among all the units of the text and the linking is in respect to all words not removed by the erasure list. Content analysis allows one to draw content links as well, except that the links are between instances of words singled out in one or the other category. What is unique to PLATEXT is that the results of the dictionary matching procedure influence the final results of the content-linking and assembly phases of the program (Chapter 3) of the analysis. In other words, the relation-indicating words in PLATEXT's dictionary are in reality operators, which, when they occur in text, modify the text structure diagram with respect to the categorized quality, in this case, relevance. Nothing prevents other connotative qualities in addition to relevance from having PLATEXT-like dictionary categories created for them, each with the power of modifying the principles and paths of the text diagram. This is where PLATEXT is related to the meta-language concepts of classificationists (88, 89, 90).

This suggests a future PLATEXT which structures text in several directions or dimensions. Each statement would



have associated with it a component indicating the level of relevance (as in the original PLATEXT), perhaps a component indicating the degree of objectivity, and perhaps a component to express the degree of personal involvement of the author. The analysis would include as many of these components as is necessary to provide the connotative background for the statement. The results of the connotative analysis might be presented to the investigator on a CRT, or the analysis could be used to relate the discourse to previously entered information in a truly "conversational" data bank. The provision of such analyses will be required if many of the visions of the next generation of information systems and networks are to be realized.



## BIBLIOGRAPHY





## BIBLIOGRAPHY

1. Silk, L. S., The Research Revolution, McGraw-Hill Book Co., Inc., New York, 1960.
2. Loosjes, Th. P., On Documentation of Scientific Literature, Archon Books, London, 1967.
3. Kochen, Manfred (Ed.), The Growth of Knowledge, John Wiley and Sons, New York, 1967.
4. Heaps, D. M. and Ingram, W. D., "Information Transfer in Canada: Position Paper", In: Proceedings of the Annual Meeting, Western Canada Chapter, American Society for Information Science, Winnipeg, September 1972 (Published by the University of British Columbia School of Librarianship, Vancouver, 1972), pp. 171-181.
5. Language and Machines, Report by the Automatic Language Processing Advisory Committee Division of Behavioral Sciences and National Research Council, Publication No. 1416, National Academy of Sciences and National Research Council, Washington, D.C., 1966.
6. Cuadra, E. (Ed.), Annual Review of Information Science and Technology, Vols. I-IV, Encyclopaedia Britannica, Inc., Chicago, 1966-1971.
7. World List of Scientific Periodicals, Butterworths, London, (Prepared by National Central Library,) 1964.
8. Union List of Scientific Serials in Canadian Libraries, National Science Library, Ottawa, 1969.
9. Belzer, J., "Education in Information Science", Journal of the American Society for Information Science, Vol. 21, No. 4, pp. 269-273, July/August 1970.
10. Weinberg, Alvin, Science, Government, and Information: The Responsibilities of the Technical Community and the Government in the Transfer of Information, President's Science Advisory Council, U.S. Government Printing Office, Washington, D.C., 1963.
11. A Science Policy for Canada, Report of the Senate Special Committee on Science Policy (the Lamontagne Report), Queen's Printer, Ottawa, 1970.



12. Abraham, J. and Adams, A. J., Developing Computer Professionals, Panel Discussion, Canadian Computer Conference, June 2 - June 3, 1972.
13. To Know and Be Known, Report of the Task Force on Government Information (the Fortier Report), Queen's Printer, Ottawa, 1969.
14. Fitzpatrick, A., (Alberta Research Council) Personal Communication, November 1972.
15. Heaps, D. M. and Cooke, G. A., "National Policies, National Networks and National Information Studies in Canada", Proceedings of the American Society for Information Science, Vol. 7, pp. 199-203, 1970.
16. FID International Congress, Budapest, September 1972, "Participation of Small and Less Industrialized Countries in International Information Exchange".
17. Consumers Association of Canada, "The power to communicate: A revolution in information-sharing". (A report on information handling in community centres in Canada, 1971/72; funded by Canadian Computer/Communications Task Force and directed by Diana Ironsides) (To be published).
18. Aitchison, J., "The Thesaurofacet: A Multipurpose Retrieval Language Tool", Journal of Documentation, Vol. 26, No. 3, pp. 187-203, September 1970.
19. Global and Long-distance Decision-making: Environmental Issues and Network Potentials, Stockholm Royal Institute of Technology and University of Stockholm, Department of Data Processing, Report, IB-ADB 72 No. 6, 1972.
20. Samuelson, Kjell, "Challenges to Theory and Methods of Systems, Cybernetics and Information Networks", Stockholm Royal Institute of Technology and University of Stockholm, Department of Data Processing, Report, IB-ADB 72 No. 9, 1972 (Paper presented at FID/TM Seminar, Budapest, September 5, 1972).
21. Coates, F. J., "Switching Languages for Indexing", Journal of Documentation, Vol. 26, No. 2, pp. 102-110, June 1970.
22. Proceedings of the International Symposium: UDC in Relation to Other Indexing Languages, Herceg Novi, Yugoslavia, 1972 (Sponsored by the Yugoslav Centre for Technical and Scientific Documentation and FID).



23. Federation International de Documentation, FID/CCC's New Policy for a Standard Reference Code (SRC), Announcement Brochure, La Haye, December 9, 1971.
24. Harris, J. L., Subject Analysis, Scarecrow Press, Metuchen, N. J., 1970.
25. Mercier, M. A., Study of the UDC and Other Indexing Languages Through Computer Manipulation of Machine Readable Data Bases, M. Sc. Thesis, University of Alberta, Edmonton, 1972.
26. Mercier, M., Cooke, G.A. and Heaps, D. M., "The Study of UDC and Other Indexing Languages Through Computer Manipulation of Machine-readable Data Bases", In: Proceedings of the International Symposium: UDC in Relation to Other Indexing Languages, Herceg Novi, Yugoslavia, 1972 (Sponsored by the Yugoslav Centre for Technical and Scientific Documentation and FID), pp. 1-24.
27. Alber, F. M., On-line Thesaurus Design for an Integrated Information System, M. Sc. Thesis, University of Alberta, Edmonton, 1972.
28. Alber, F. and Heaps, D. M., "Classifying, Indexing, and Searching Resource Management Information Via an On-line Thesaurus", In: Proceedings of the Annual Meeting, Western Canada Chapter, American Society for Information Science, Banff, 1971 (Published by Information Systems, University of Calgary), pp. 107-116.
29. Dobay, S. R. and Heaps, D. M., "Machine-Readable Subject Authority Lists and Thesauri as Aids in Classification and Concept Analysis", In: Proceedings of the Annual Meeting, Western Canada Chapter, American Society for Information Science, Winnipeg, September 1972 (Published by the University of British Columbia School of Librarianship, Vancouver, 1972), pp. 159-170.
30. Heaps, D. M. and Ingram, W. D., "Computer Recognition and Graphical Reproduction of Patterns in Scientific and Technical Style". Proceedings of the American Society for Information Science, Vol. 8, pp. 257-261, 1971.
31. Chomsky, N., "A Review of B. F. Skinner's Verbal Behavior", Language, Vol. 35, no. 1, pp. 26-58, 1959.
32. Chomsky, Noam, "Current Issues in Linguistic Theory", The Structure of Language, Katz and Fodor (Eds.), Prentice-Hall, Inc. Englewood Cliffs, N.J., 1964. pp. 50-118.





33. Thompson, Frederick, "English for the Computer", Proceedings of the Fall Joint Computer Conference, 1966, pp. 349-356.
34. Fairthorne, R. A., Temporal Structure in Bibliographical Classification, In: Proceedings of the First Ottawa Conference on the Conceptual Basis of the Classification of Knowledge, University of Ottawa, 1971 (Preprint).
35. Coblans, H., "Words and Documents", ASLIB Proceedings, Vol. 23, No. 7, pp. 337-350, July 1971.
36. Fairthorne, R. A., Seminar on Education for Information Science, Sponsored by the American Library Association and the American Society for Information Science, Denver, September 1971 (Discussant).
37. Thompson, Frederick B., "The Organization is the Information", American Documentation, Vol. 19, pp. 305-308, 1968.
38. The Christensen Rhetoric Program, The Sentence and The Paragraph, (Teacher's Manual), Harper & Row, New York, 1968.
39. Kucera, H., and Francis, W. N., Computational Analysis of Present Day American English, Brown University Press, Providence, Rhode Island, 1967.
40. Leedham, Charles, "The Chip -- Modern Marvel of Electronics", The New York Times Magazine, September 19, 1965.
41. Asimov, Isaac, The Genetic Code, Orion Press, New York, 1962.
42. Huxley, Julian, Evolution: The Modern Synthesis, Harper & Row, New York, 1942.
43. Clowes, Royston, The Structure of Life, Penguin Books Ltd., Middlesex, England, 1967.
44. Edgar, R. S. and Epstein, R. H., "The Genetics of a Bacterial Virus", Scientific American, February 1965.
45. Lowenstein, C. D. and Anderson, V. D., "Quick characterization of the directional response of point arrays", Jour. Acoustical Soc. of America, Vol. 43, pp. 32-36, 1968.
46. Hutchison, Gordon, "Round-up", Canadian Electronics Engineering, Vol. 13, No. 3, p. 80, March 1969.





47. "Getting in on Ocean Decade", Canadian Research & Development, Vol. 2, pp. 38-39, March/April 1969.
48. Galko, John A., "Thick-Film Hybrids", Canadian Electronics Engineering, Vol. 13, No. 3, pp. 49-52, March 1969.
49. Hoyle, Fred, "Science, Society, Action, Reaction", Physics Today, pp. 148-150, April 1968.
50. Cameron, A., Frank, R. and Leyerle, J. (Eds), Computers and Old English Concordances, University of Toronto Press, Toronto, 1970.
51. "Big Business Gets Word: R & D Spending Inadequate", The Financial Post, June 7, 1969.
52. A Programmer's Introduction to the IBM System 360/Architecture, Instructions, and Assembler Language, No. C20-1618, IBM, 1966.
53. Harris, Z. S., "Discourse Analysis", Language, Vol. 28 No. 4, pp. 1-28, 1952.
54. Jacobson, S. N., "A Modifiable Routine for Connecting Related Sentences of English Text". In: Computation in Linguistics, University of Indiana Press, 1966, pp. 284-311.
55. Pritchard, A., "Statistical Bibliography or Bibliometrics?", Documentation Note; Journal of Documentation, Vol. 25, No. 4, pp. 348-349, December, 1969.
56. Fairthorne, R. A., "Empirical Hyperbolic Distributions (Bradford-Zipf-Mandelbrot) for Bibliometric Description and Prediction", Journal of Documentation, Vol. 25, No. 4, pp. 319-342, December 1969.
57. Zipf, George Kingsley, Human Behavior and the Principle of Least Effort, Addison-Wesley Publishing Co., Cambridge, Mass., 1949.
58. Estoup, J. B., Gammes Stenographiques, 4th Edition, 1916.
59. Mandelbrot, B., "A Note on a Class of Skew Distribution Functions", Information and Control, Vol. 2, pp. 90-99, 1959.
60. Mandelbrot, B., "Final Note on a Class of Skew Distribution Functions", Information and Control, Vol. 4, pp. 198-216, 1961.



61. Mandelbrot, B., "Post Scriptum to 'Final Note'",  
Information and Control, Vol. 4, pp. 300-304, 1961.
62. Miller, G. A., "Some Effects of Intermittent Science",  
American Journal of Psychology, Vol. 70, pp. 311-314, 1957.
63. Simon, H. A., "On a Class of Skew Distribution Functions", Biometrika, Vol. 42, pp. 425-440, 1955.
64. Simon, H. A., "Some Further Notes on a Class of Skew Distribution Functions", Information and Control, Vol. 2, pp. 80-88, 1960.
65. Simon, H. A., "Reply to 'Final Note' by Benoit Mandelbrot", Information and Control, Vol. 4, pp. 217-223, 1961.
66. Herdan, Gustav, Language as Choice and Chance, P. Noordhoff, Groningen, 1956.
67. Herdan, Gustav, The Advanced Theory of Language as Choice and Chance, Springer-Verlag, New York, 1966.
68. Muller, Charles, "Lexical Distribution Reconsidered: The Waring-Herdan Formula", In: Statistics and Style, (Mathematical Linguistics and Automatic Language Processing Series, No. 6), Elsevier, New York, 1969, pp. 42-56.
69. Mandelbrot, B., "On the Theory of Word Frequencies and On Related Markovian Models of Discourse", Proceedings of Symposia in Applied Mathematics: Structure of Language and its Mathematical Aspects, pp. 190-219, 1961.
70. Mosteller, F. and Wallace, D. L., "Inference in an Authorship Problem", Journal of the American Statistical Association, Vol. 58, pp. 275-309, 1963.
71. Ellegard, A., A Statistical Method of Determining Authorship, University of Gothenburg, 1962.
72. Sedelow, S. Y. and Bobrow, D. G., "A LISP Program for Use in Stylistic Analysis", TM-1753, System Development Corp., Santa Monica, Calif., February 17, 1964.
73. Sedelow, S. Y., Sedelow, W. A. Jr. and Ruggles, T. L., "Some Parameters for Computational Stylistics: Computer Aids to the Use of Traditional Categories in Stylistic Analysis", Proceedings of the IBM Literary Data Processing Conference, September 9-11, 1964, pp. 211-229.





74. Stone, P. J., Dunphy, D. C., Smith, M. S. and Ogilvie, D. G., The General Inquirer: A Computer Approach to Content Analysis, M. I. T. Press, Cambridge, Mass., 1966.
75. Psathas, George, "The General Inquirer: Useful or Not?", Computers and the Humanities, Vol. 3, No. 3, pp. 163-174, January 1969.
76. Luhn, H. P., "The Automatic Creation of Literature Abstracts", IBM Journal of Research and Development, Vol. 2, No. 2, April, 1958.
77. Petrarca, A. E., Laitimer, S. V. and Lay, Wm., "Use of the Double KWIC Coordinate Indexing Technique for Chemical Line Notations", Ohio State University, Computer and Information Science Research Center, Technical Report 70-9, 1970.
78. Landry, B. C. and Rush, J. E., "Towards a Theory of Indexing", Proceedings of the American Society for Information Science, Vol. 5, pp. 59-63, 1968.
79. Landry, B. C. and Rush, J. E., "Towards a Theory of Indexing II", Journal of the American Society for Information Science, Vol. 21, No. 5, pp. 358-367, September/October 1970.
80. Rush, J. E., Salvador, R. and Zamora, A., "Automatic Abstracting and Indexing. II Production of Indicative Abstracts by Application of Contextual Inference and Syntactic Coherency Criteria", Journal of the American Society for Information Science, Vol. 22, No. 4, pp. 260-274, 1971.
81. Hays, David G., Introduction to Computational Linguistics, (Mathematical Linguistics and Automatic Language Processing, No. 2), Elsevier, New York, 1967.
82. Borko, H., Automated Language Processing, John Wiley and Sons, New York, 1967.
83. Lowenthal, Leo, Literature, Popular Culture, and Society, Prentice-Hall, Inc., Englewood Cliffs, N. J., 1961.
84. Harrison, James (Ed.), Scientist as Writers, M.I.T. Press, Cambridge, Mass., 1965.
85. Dolezel, L. and Bailey, R. W. (Eds.), Statistics and Style, (Mathematical Linguistics and Automatic Language Processing Series, No. 6), Elsevier, New York, 1969.
86. Dictionary of Canadian English, W. J. Gage, Toronto, 1967.





87. Treleaven, R., Abbreviation of English Words to Standard Length for Computer Processing, M. Sc. Thesis, University of Alberta, Edmonton, 1970.
88. Gardin, Jean-Claude, "On Some Reciprocal Requirements of Linguistics and Information Techniques". In: Information in the Language Sciences, (Mathematical Linguistics and Automatic Language Processing Series, No. 5), Elsevier, New York, 1968, pp. 95-103.
89. Freeman, Robert R., "Actual and Potential Role of the Universal Decimal Classification in Language-Sciences Documentation", In: Information in the Language Sciences, (Mathematical Linguistics and Automatic Language Processing Series, No. 5), Elsevier, New York, 1968, pp. 149-163.
90. Austin, D., "Trends Toward a Compatible General System", In Classification in the 1970's, Bingley, London, 1972, pp. 213-248.



## APPENDIX A



## APPENDIX A

### Dictionary of Relation Indicators (September, 1972)

Entry	Type	Weight
about the same time	coordinate	1
also	subordinate	2
although	subordinate	1
as a result	subordinate	2
as if	subordinate	2
at the present time	coordinate	1
by this time	coordinate	2
but	coordinate	1
certainly	coordinate	1
consider	superordinate	2
example	subordinate	3
finally	coordinate	1
fortunately	subordinate	1
further	coordinate	1
furthermore	coordinate	1
however	coordinate	1
if	coordinate	1
indeed	coordinate	1
instance	subordinate	2
instead	subordinate	2
in addition	coordinate	2
in consequence	superordinate	1



Dictionary of Relation Indicators (continued)

Entry	Type	Weight
in contrast	coordinate	3
in due course	coordinate	1
in general	superordinate	2
in that case	subordinate	2
in the case	subordinate	2
in this case	subordinate	2
its	subordinate	2
i.e.	subordinate	3
just as	superordinate	1
latter	subordinate	2
merely	subordinate	2
namely	superordinate	1
neither	coordinate	1
nevertheless	superordinate	1
nor	coordinate	1
occasionally	subordinate	1
of course	subordinate	2
only	subordinate	1
on the contrary	coordinate	1
on the other hand	coordinate	2
perhaps	subordinate	2
principal	superordinate	1
principle	superordinate	1
probably	subordinate	1
rather	coordinate	1





Dictionary of Relation Indicators (continued)

Entry	Type	Weight
second	subordinate	1
significant	superordinate	1
similar	coordinate	1
since	subordinate	2
some of these	subordinate	2
sometimes	subordinate	2
somewhat	subordinate	1
still	coordinate	1
surely	coordinate	1
that is	subordinate	2
that is to say	subordinate	2
then	coordinate	1
therefore	superordinate	3
third	subordinate	1
this	coordinate	1
thus	superordinate	2
to sum up	superordinate	3
unfortunately	subordinate	1
unique	superordinate	1
up till now	subordinate	2
when	subordinate	1
whether	subordinate	1
yet	coordinate	1



## APPENDIX B



## APPENDIX B

### Erasure list

the	no	your	go
of	if	way	came
and	out	well	right
to	so	down	used
a	said	should	take
in	what	because	three
that	up	each	states
is	its	just	himself
was	about	those	few
he	into	people	house
for	than	Mr	use
it	them	how	during
with	can	too	without
as	only	little	again
his	other	state	place
on	new	good	American
be	some	very	around
at	could	make	however
by	time	world	home
I	these	still	small
this	two	own	found
had	may	see	Mrs
not	then	men	thought
are	do	work	went
but	first	long	say
from	any	get	part
or	my	here	once
have	now	between	general
an	such	both	high
they	like	life	upon
which	over	under	every
you	man	never	don't
were	me	day	does
her	even	same	got
all	most	another	united
she	made	know	left
there	after	while	number
would	also	last	course
their	did	might	war
we	many	us	until
him	before	great	always
been	must	old	away
has	through	year	something
when	back	off	fact
who	years	come	though
will	where	since	water
more	much	against	less
one	our	being	school





Erasure List (continued)

public	possible	certain	week
put	rather	kind	car
think	second	problem	field
almost	face	began	word
hand	per	different	words
enough	among	door	already
far	form	thus	themselves
took	important	help	information
head	often	sense	I'm
yet	things	means	tell
government	looked	whole	college
system	early	matter	shall
better	white	perhaps	together
set	case	itself	money
told	John	it's	period
nothing	become	York	held
night	large	times	keep
end	big	human	sure
why	need	law	probably
called	four	line	free
didn't	within	above	real
eyes	felt	name	seems
find	along	example	behind
going	children	action	cannot
look	saw	company	miss
asked	best	hands	political
later	church	local	air
knew	ever	show	question
point	least	five	making
next	power	history	office
program	development	whether	brought
city	light	gave	whose
business	thing	either	special
give	seemed	today	heard
group	family	act	major
toward	interest	feet	problems
young	want	across	ago
days	members	past	became
let	mind	quite	federal
room	country	taken	moment
president	area	anything	study
side	others	having	available
social	done	seen	known
given	turned	death	result
present	although	body	street
several	open	experience	economic
order	God	half	boy
national	service	really	position



Erasure List (continued)

reason	child	personal
change	effect	process
south	level	situation
board	students	alone
individual	military	English
job	run	gone
society	short	idea
areas	stood	increase
west	town	nor
close	morning	schools
turn	total	women
love	outside	
community	figure	
true	rate	
court	art	
force	century	
full	class	
cost	north	
seem	usually	
am	Washington	
wife	leave	
age	plan	
future	therefore	
voice	evidence	
wanted	million	
department	sound	
center	top	
woman	black	
common	hard	
control	strong	
necessary	various	
policy	believe	
following	play	
front	says	
sometimes	surface	
girl	type	
six	value	
clear	mean	
further	soon	
land	lines	
able	modern	
feel	near	
mother	peace	
music	table	
party	red	
provide	road	
education	tax	
university	minutes	



## APPENDIX C



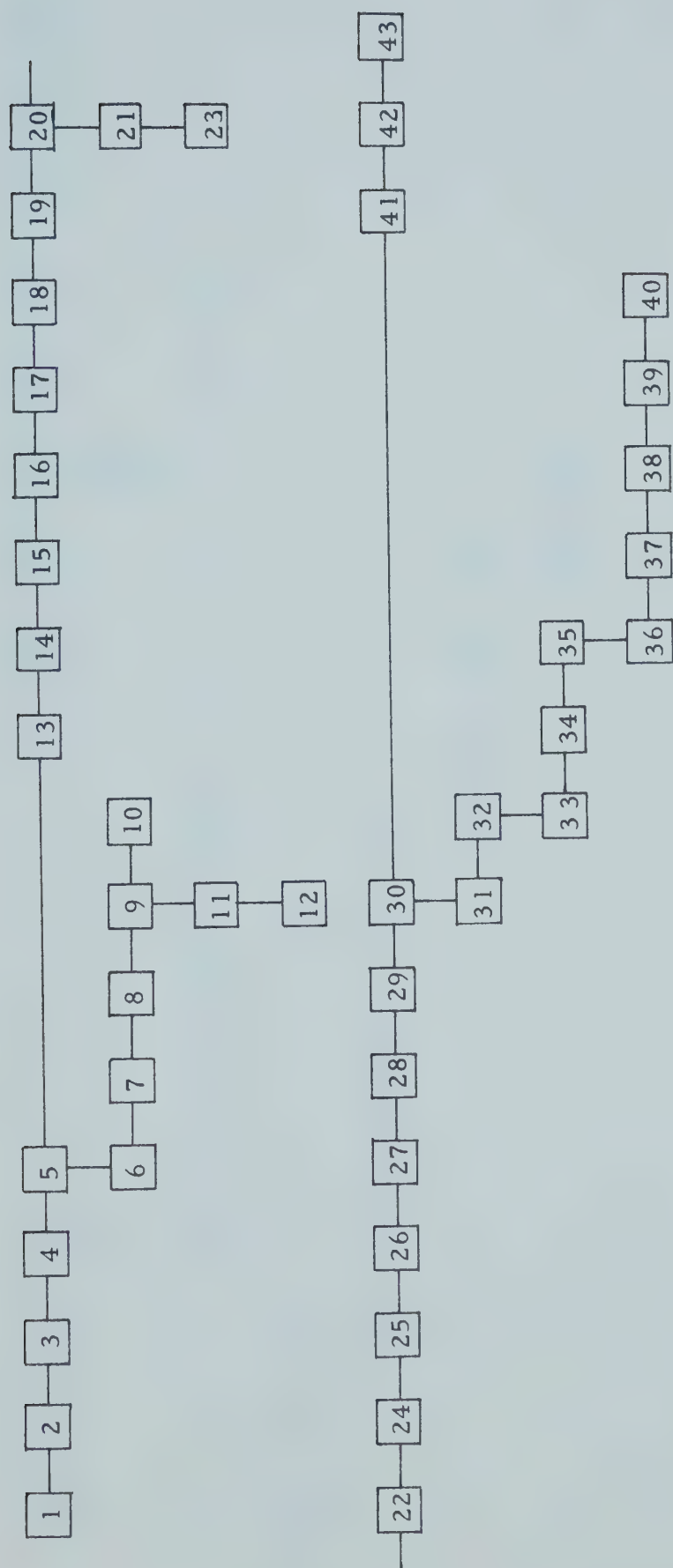


Figure C-1: Article on Hydrophone Arrays





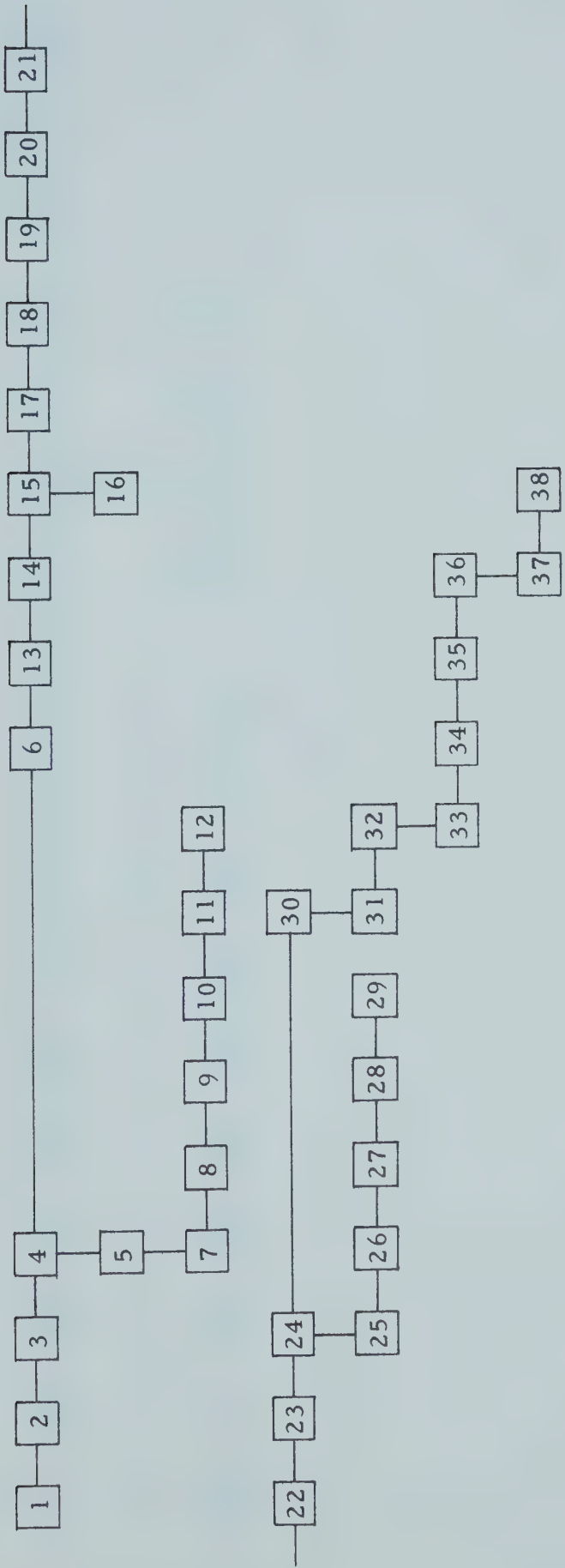


Figure C-2: Integrated Circuits (part 1), CEE



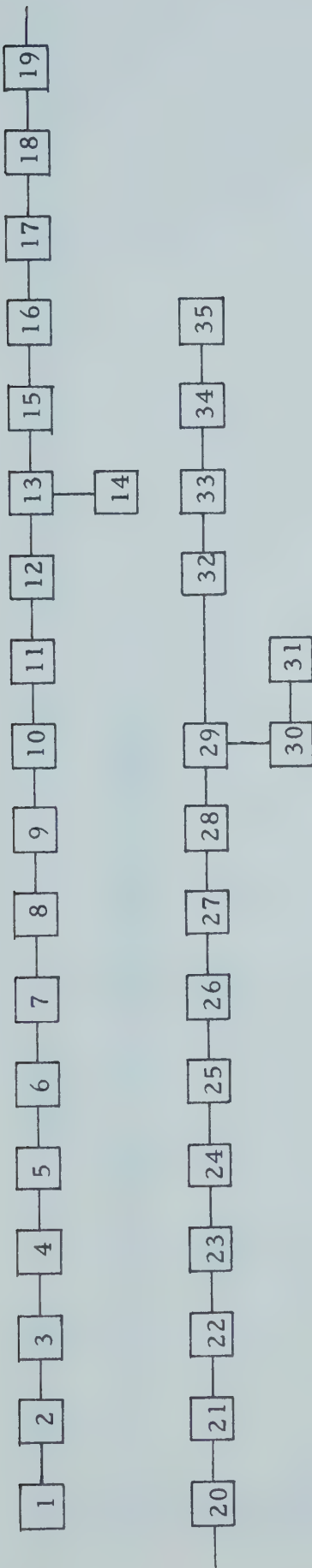
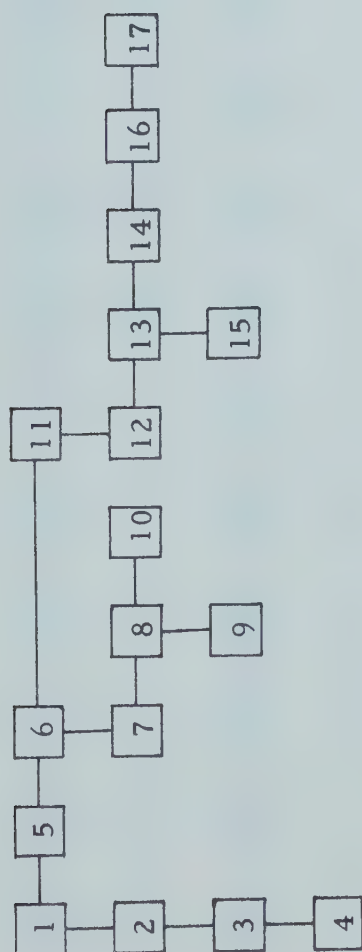


Figure C-3: Integrated Circuits (part 2), CEE









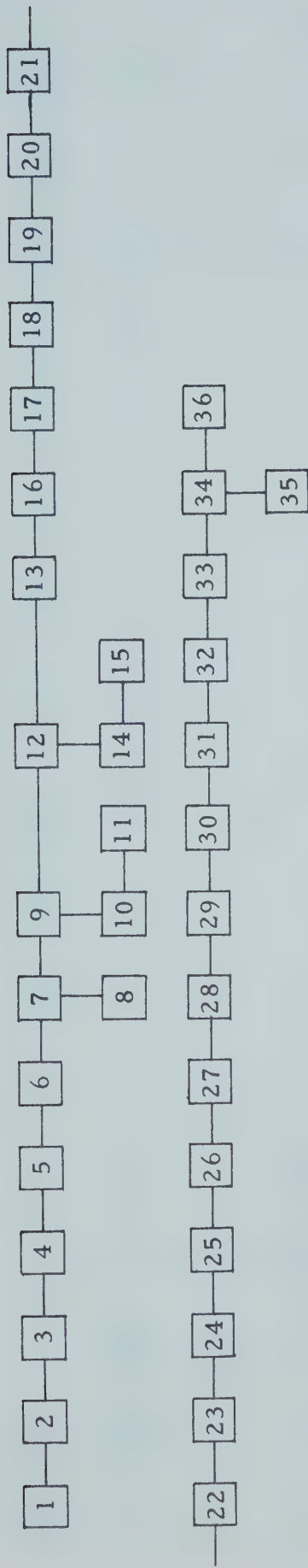


Figure C-5: Article by F. Hoyle



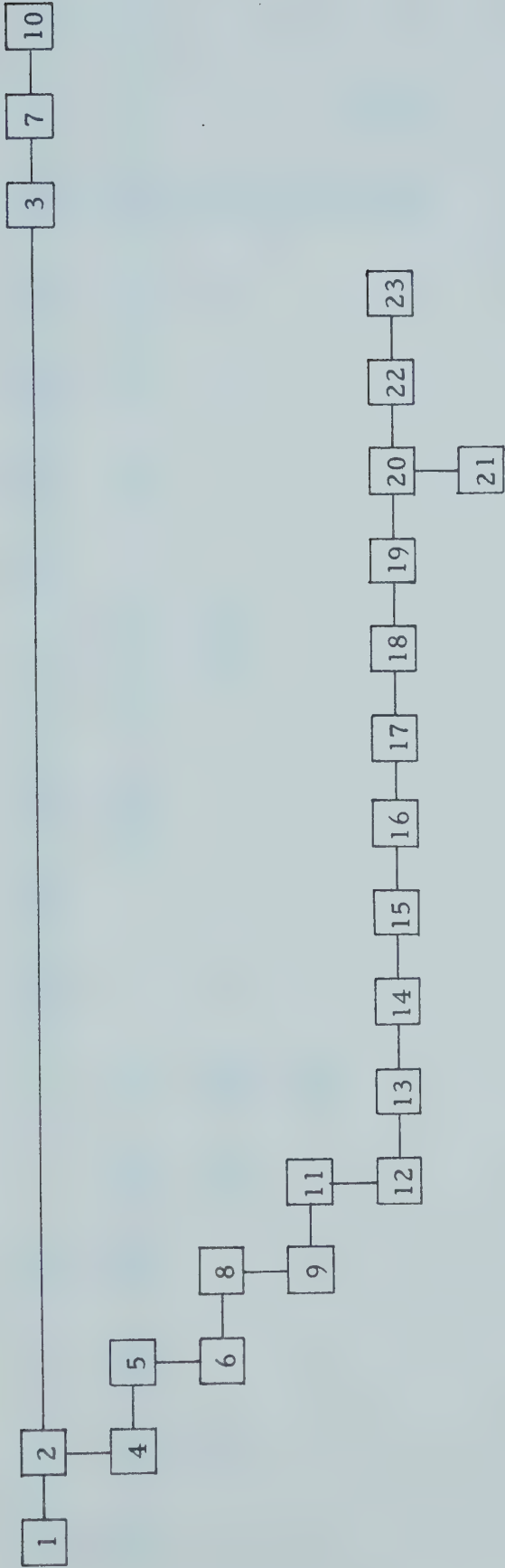


Figure C-6: Paper by Heaps and Ingram



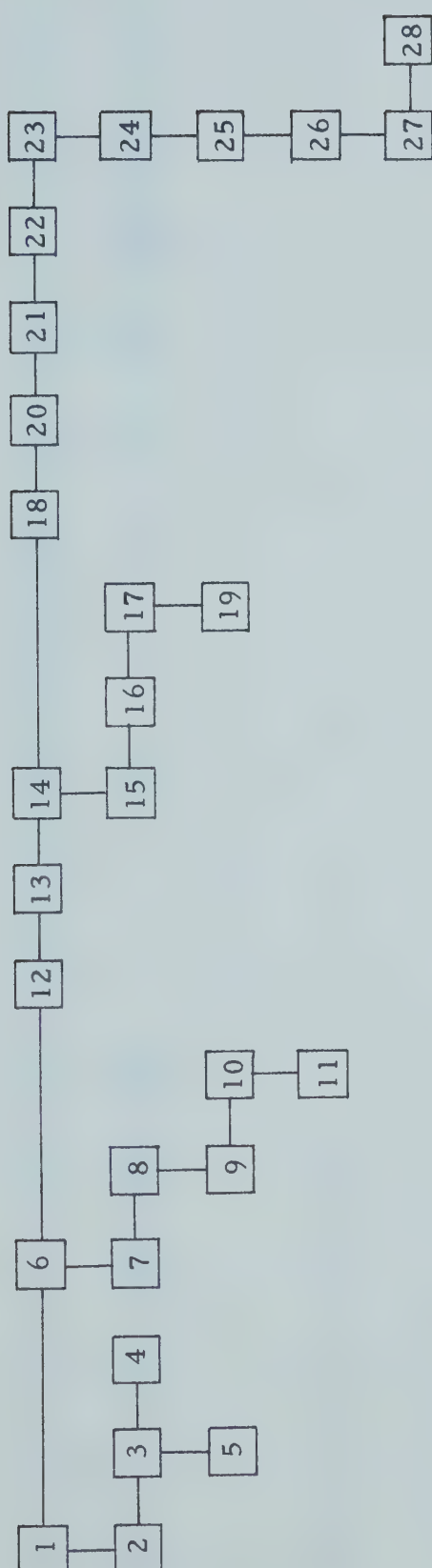


Figure C-7: Textbook on Genetics



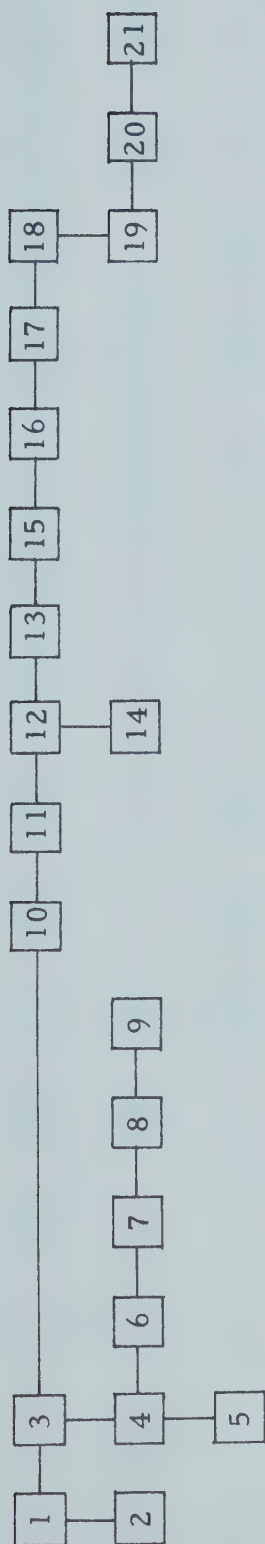


Figure C-8: Article from Scientific American





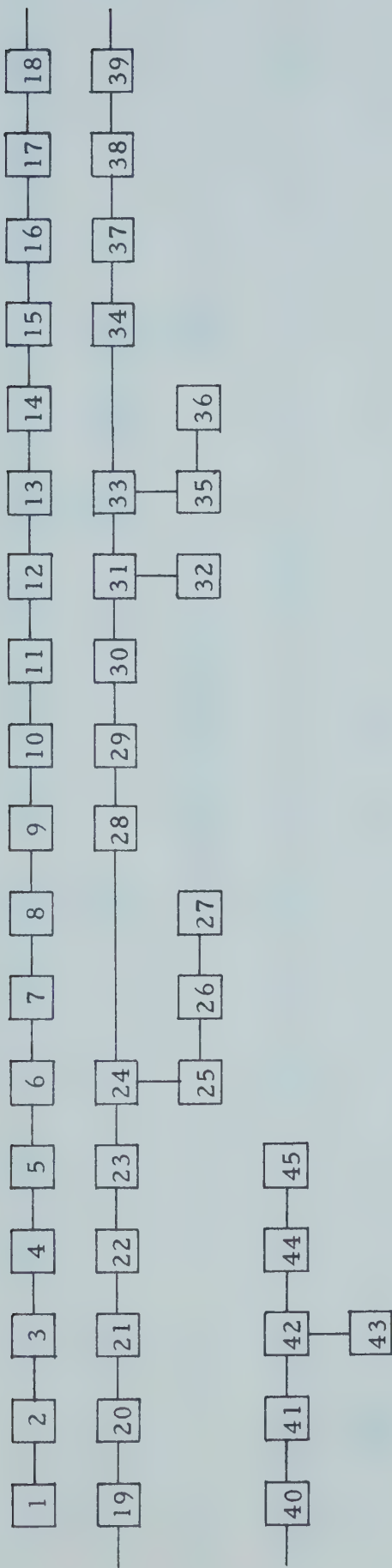


Figure C-9: Seminar on Old English Concordances



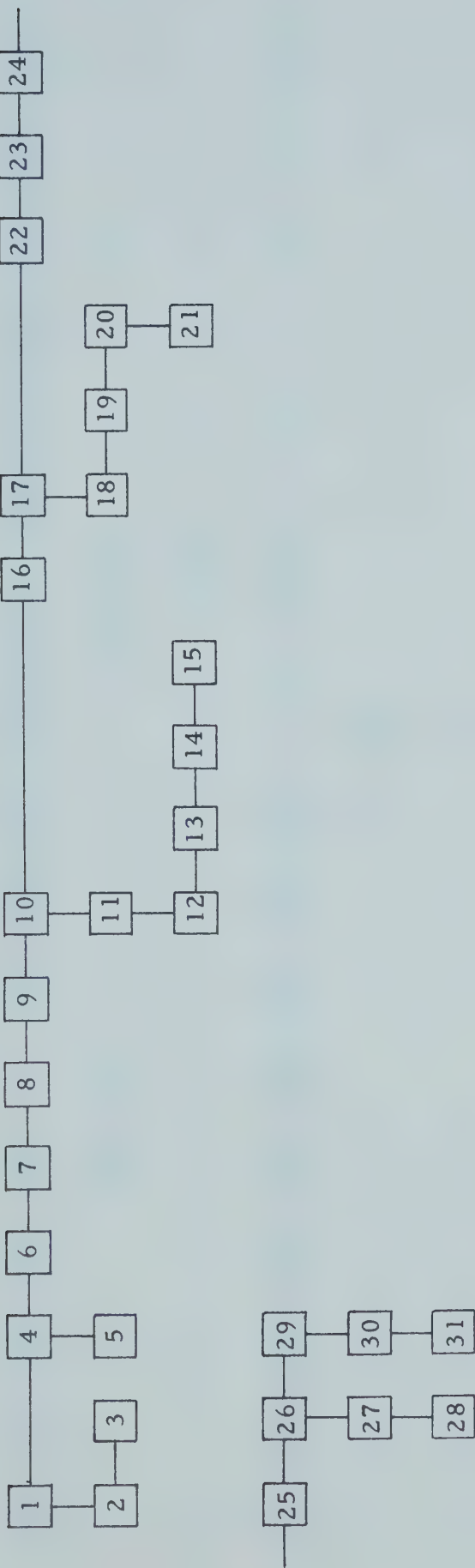


Figure C-10: Book by I. Asimov



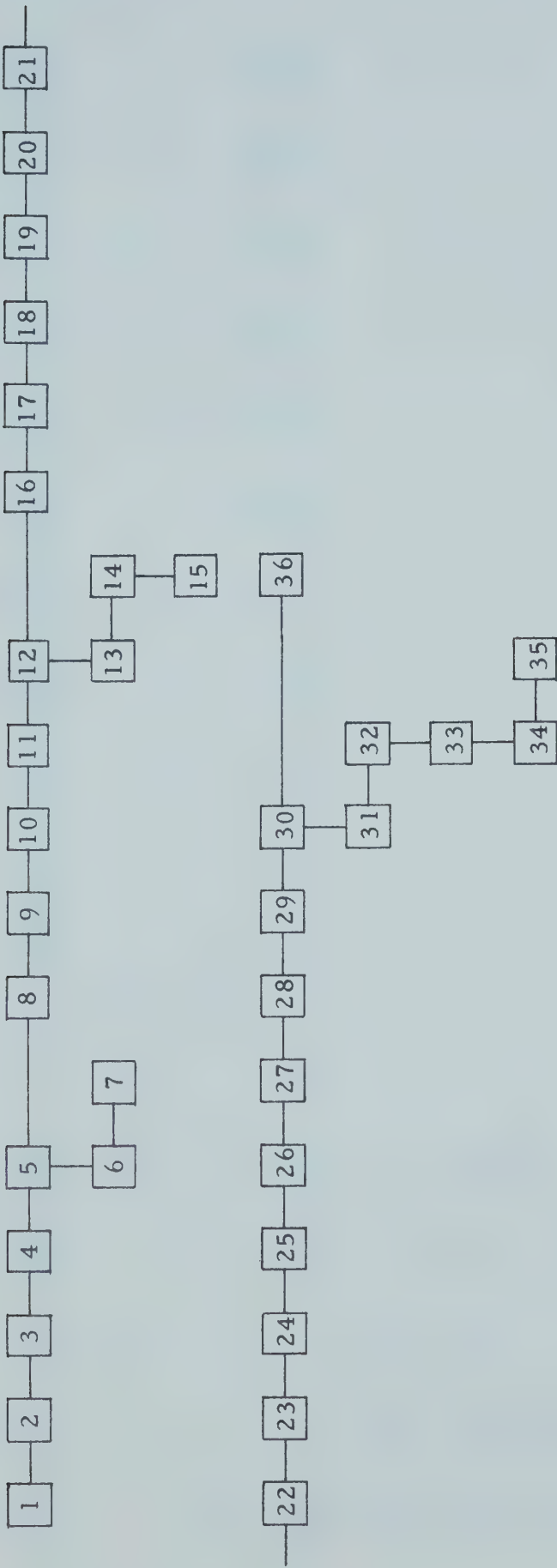


Figure C-11: Editorial, Canadian R & D





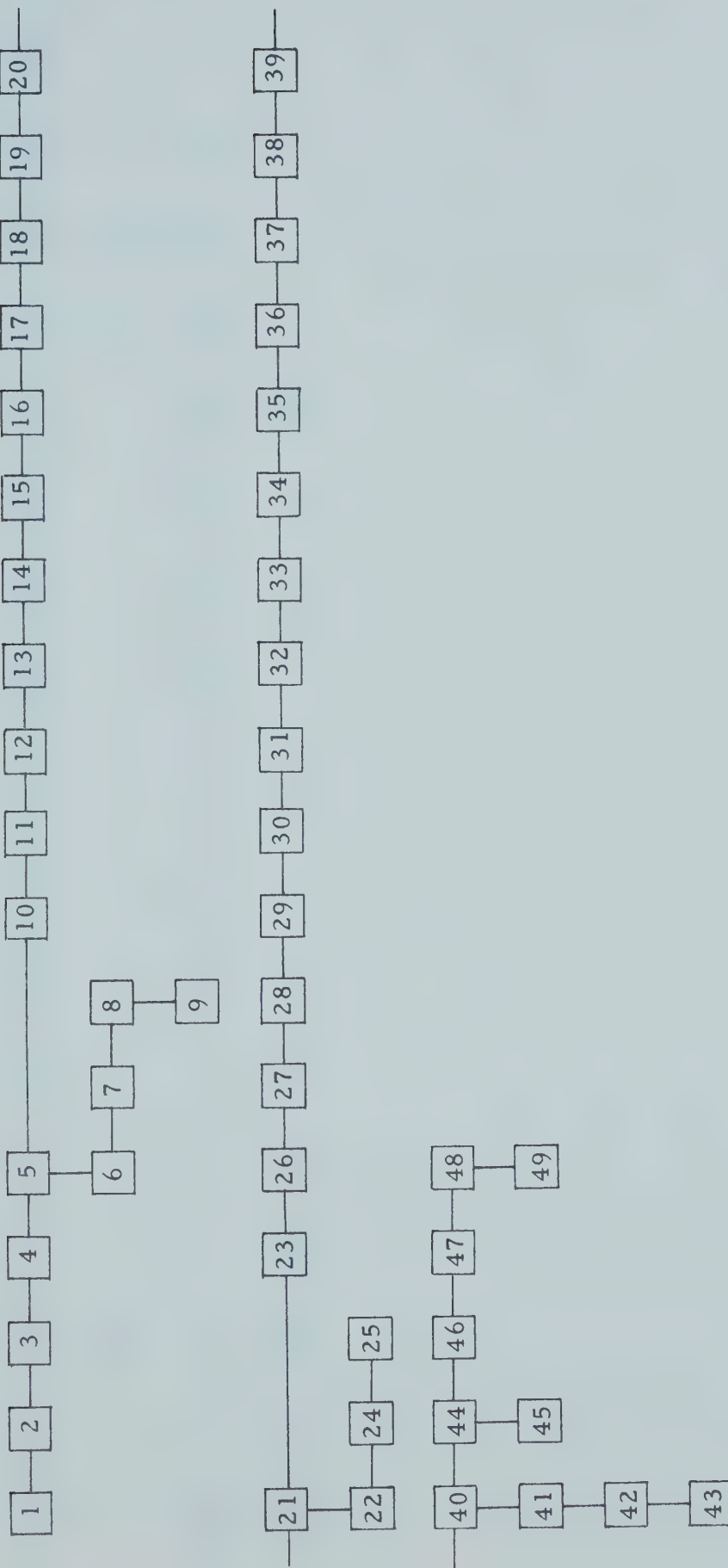


Figure C-12: Newspaper Supplement Article "The Chip"



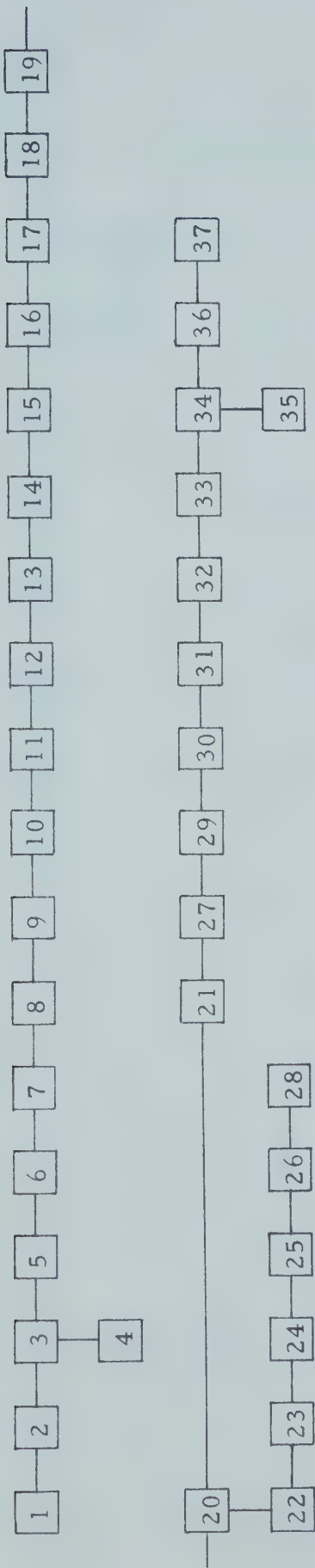


Figure C-13: Article from Financial Post





Figure C-14: Article from Canadian R & D



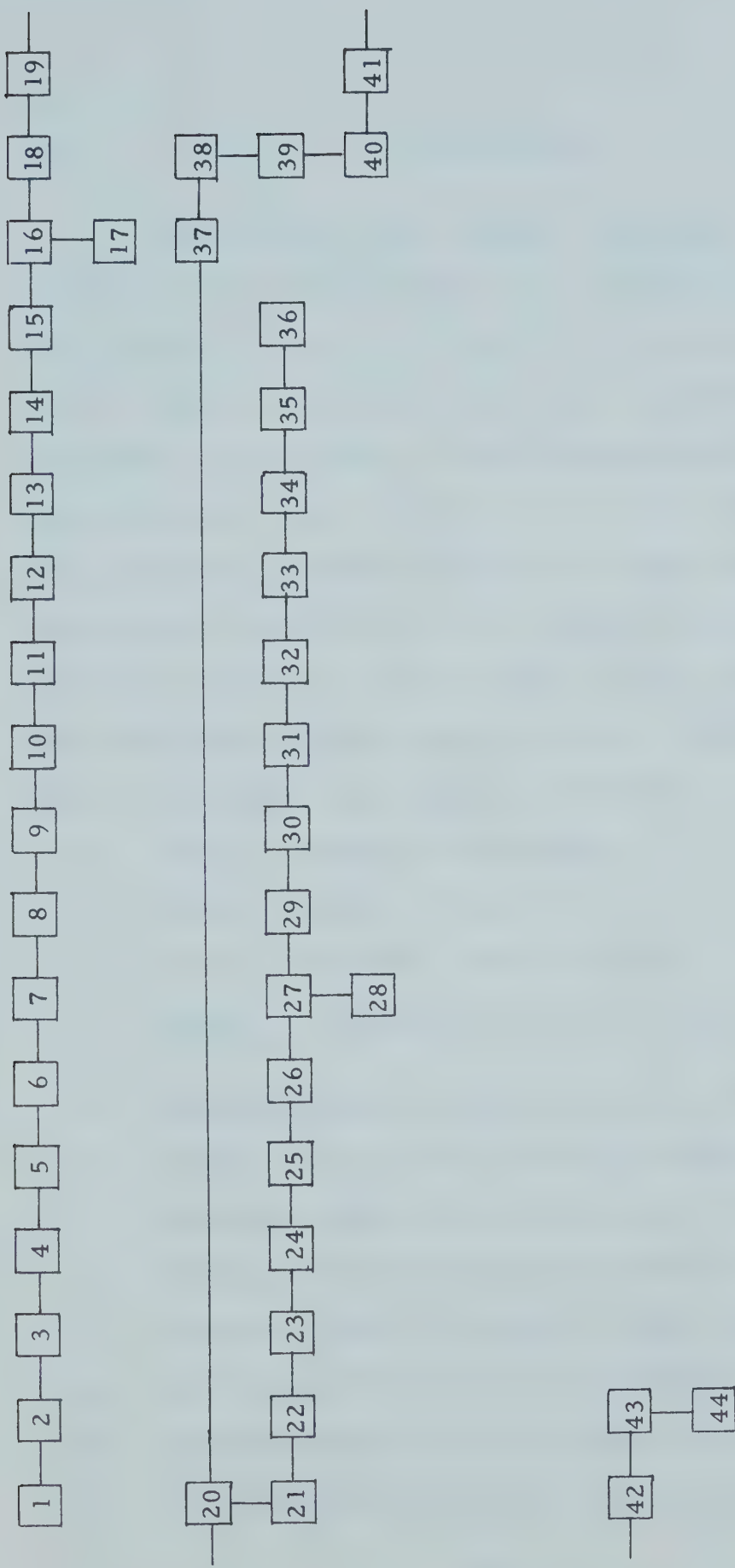


Figure C-15: IBM Manual





## APPENDIX D

The fifteen text samples, discussed in some detail in Section 4.2 of the thesis, are referenced in the bibliography as Nos. 30 and 40 through 52; these source references may be compared with the figures in Appendix C by those interested in analyses of the type performed on Sample 10 (Reference No. 41) in Section 4.1 of the thesis. The computer program produced printouts of each text sample containing, in addition to the sentence-by-sentence text listing, a listing of the content words encountered, the HORZ and VERT matrices established by PLATEXT for the sample, and the following text sample statistics:

1. Overall number of sentences.
2. Overall number of words.
3. Average sentence length overall.
4. Number of sentences on first (top) level of text diagram.
5. Average sentence length of sentences on first level.
6. Number of dictionary entries encountered.
7. Average number of dictionary entries per sentence.
8. Number of content words encountered.
9. Average number of content words per sentence.
10. Dictionary entries as a percentage of total words.
11. Content words as a percentage of total words.
12. High frequency words as a percentage of total words.



These computer printouts have not been bound in with the thesis. Information regarding the obtaining of copies of the printouts may be had from the author or the thesis supervisor. It is hoped to issue the printouts of samples and of the source listings (see Appendix F) in microfiche in connection with a separate publication.



## APPENDIX E





## APPENDIX E

### Sample Text, after Jacobson

1. But in Germany, in weird post-war Germany, he seemed snuffed out again.
2. The air was so cold and vacant, all feeling seemed to have gone out of the country.
3. Emotion, even sentiment, was numbed quite dead, as in a frost-bitten limb.
4. And if the sentiment were numbed out of him, he was truly dead.
5. "I'm most frightfully glad you've come, Kathy," he said.
6. "I could hardly have held out another day here, without you.
7. I feel you're the only thing on earth that remains real."
8. "You don't seem very real to me," she said.
9. "I'm not real!
10. I'm not! -- not when I'm alone.
11. But when I'm with you I'm the most real man alive.
12. I know it!"
13. This was the sort of thing that had fetched her in the past, thrilled her through and through in her womanly conceit, even made her fall in love with the little creature who could so generously admit such pertinent truths.
14. So different from the lordly Alan, who expected a woman to bow down to him!
15. Now, however some of the coldness of numbed Germany seemed to have got into her breast too.
16. She felt a cruel derision of the whimpering little beast who claimed reality only through a woman.
17. She did not answer him, but looked out at the snow falling between her and the dark trees.
18. Another world!



Sample Text, after Jacobson (continued)

19. When the snow left off, how bristling and ghostly the cold fir-trees looked, tall, conical creatures crowding darkly and half-whitened with snow!
20. So tall, so wolfish!
21. Phillip shivered and looked yellower.
22. There was shortage of fuel, shortage of food, shortage of everything.
23. He wanted Katherine to go to Paris with him.
24. But she would stay at least two weeks near her people.
25. The shortage she would put up with.
26. She saw at evening the string of decent townsfolk waiting in the dark--the town was not half-lighted--to fill their hot-water bottles at the hot spring outside the Kurhaus, silent, spectral, unable to afford fire to heat their own water.
27. And she felt quite cold about Phillip's shivering.
28. Let him shiver.

D. H. Lawrence, 'The Border Line', in The Woman Who Rode Away and Other Stories, Berkley Publishing Corporation, New York City, January 1962, pp. 84, 85.



## APPENDIX F

As stated in the thesis, PLATEXT was written in FORTRAN IV and was implemented on an IBM 360/67 under, at various times, OS, CP/CMS, and MTS. The program itself, stripped of lists, dictionary, comments, etc. consisted of 845 FORTRAN statements in the main program and subroutines. The computer printout of the source listings is not bound in the thesis. The detailed flowcharts in the body of the thesis adequately describe the algorithms employed in the analysis of the sample texts and the synthesis by computer of the text diagrams. As stated in Appendix D, it is hoped to issue the printouts of text samples and source listings of the program PLATEXT in microfiche in connection with a separate publication. In the meantime, those interested in copies of the source listing printout should contact the author or the thesis supervisor.





















**B30047**